Explainability & Common Robustness

马兴军，复旦大学 计算机学院

Recap: week 1

1. What is Machine Learning

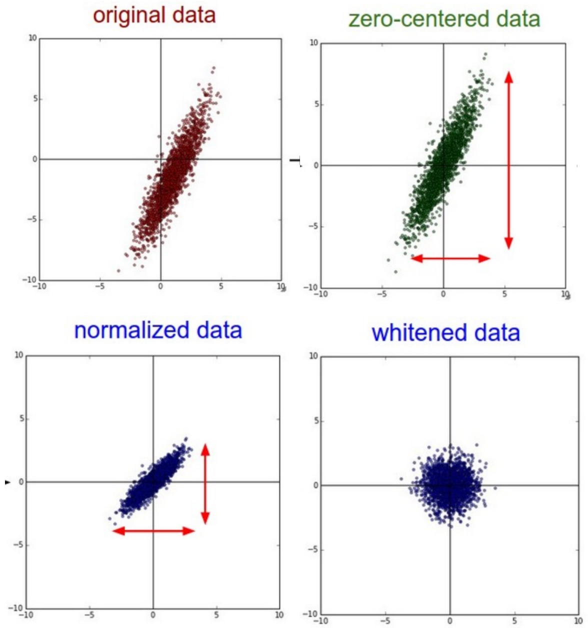
2. Machine Learning Paradigms

3. Loss Functions

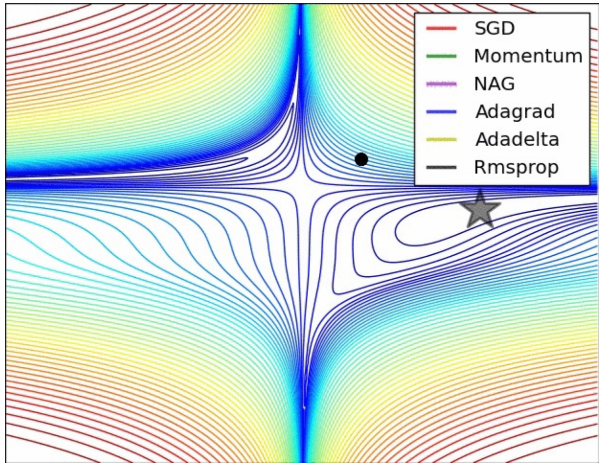
4. Optimization Methods

Machine Learning Pipeline

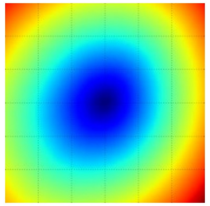
setup the input



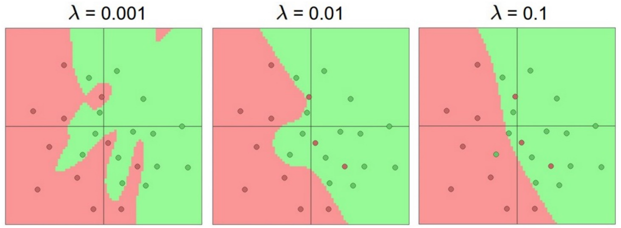
setup the optimiser



setup the loss



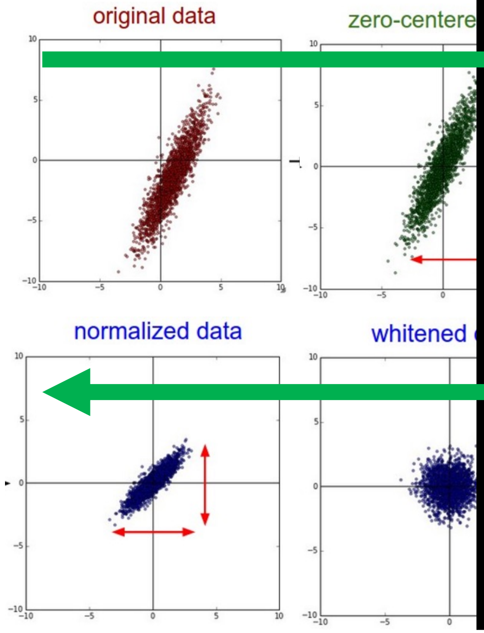
landscape of a loss function, it varies w.r.t. data, the function itself



regularization makes decision region smoother

Machine Learning Pipeline

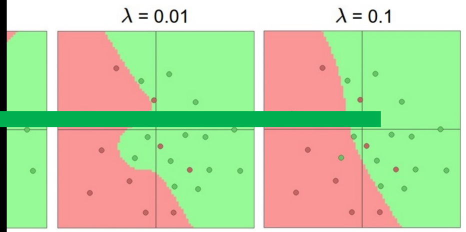
setup the input



Model ?

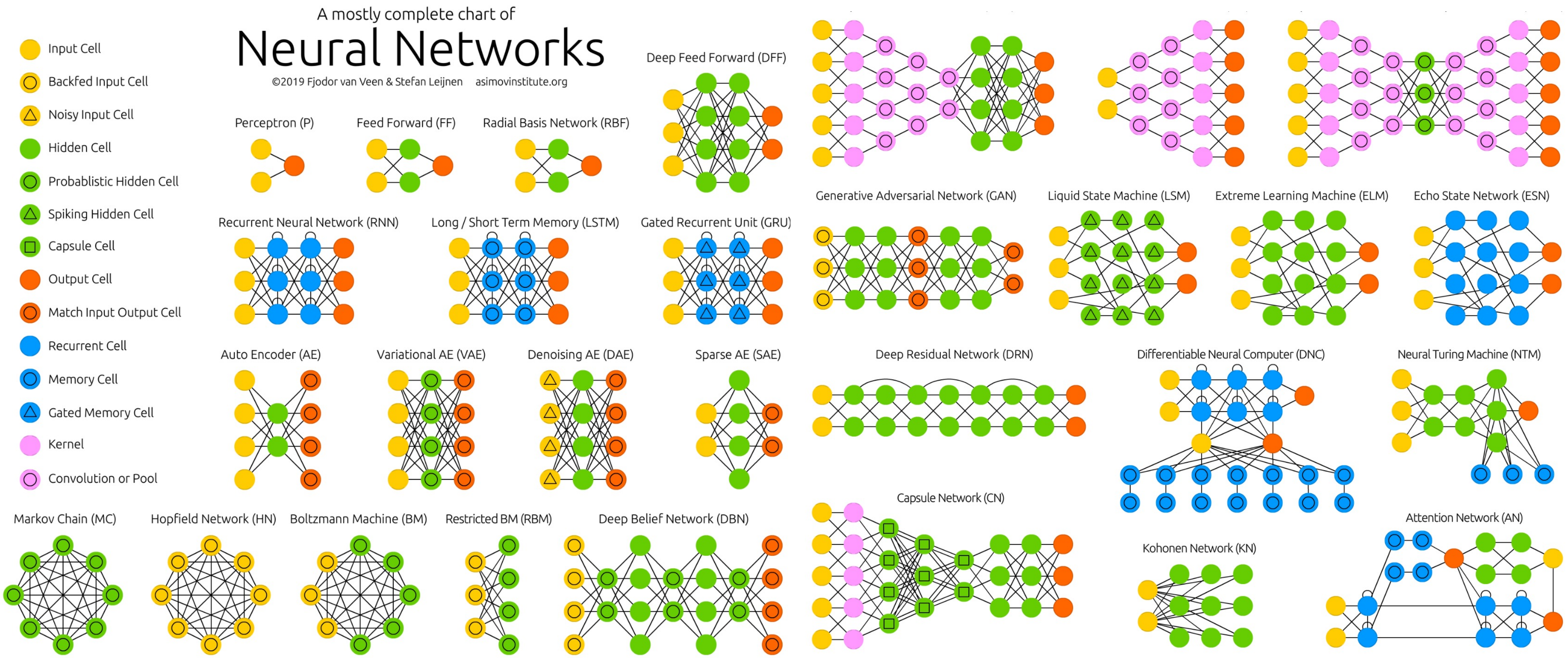
tup the loss

landscape of a loss function, it varies w.r.t. data, the function itself



ization makes decision smoother

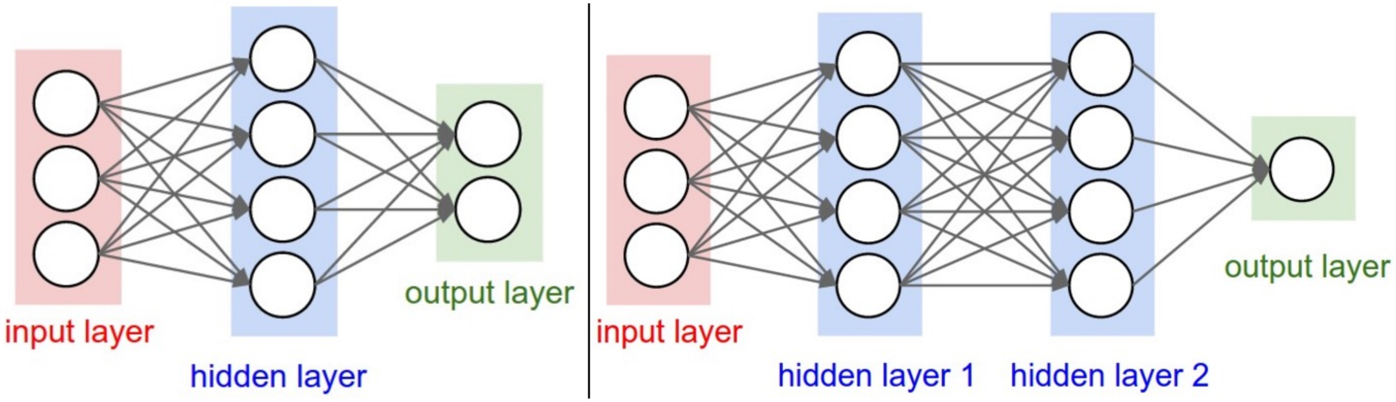
Deep Neural Networks

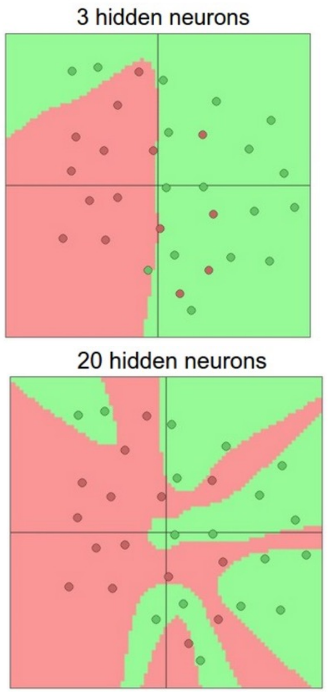


https://www.asimovinstitute.org/neural-network-zoo/; https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

Feed-Forward Neural Networks

Feed-Forward Neural Networks (FNN)
Fully Connected Neural Networks (FCN)
Multilayer Perceptron (MLP)

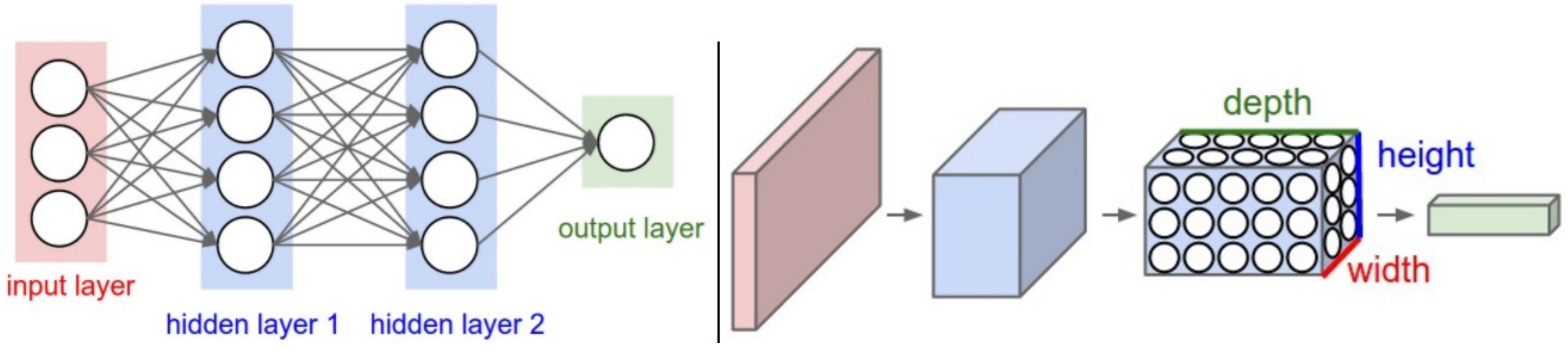




- The **simplest** neural network
- **Fully-connected** between layers
- For data that has NO temporal or spatial order

http://cs231n.stanford.edu/

Convolutional Neural Networks



Neurons in one flat layer

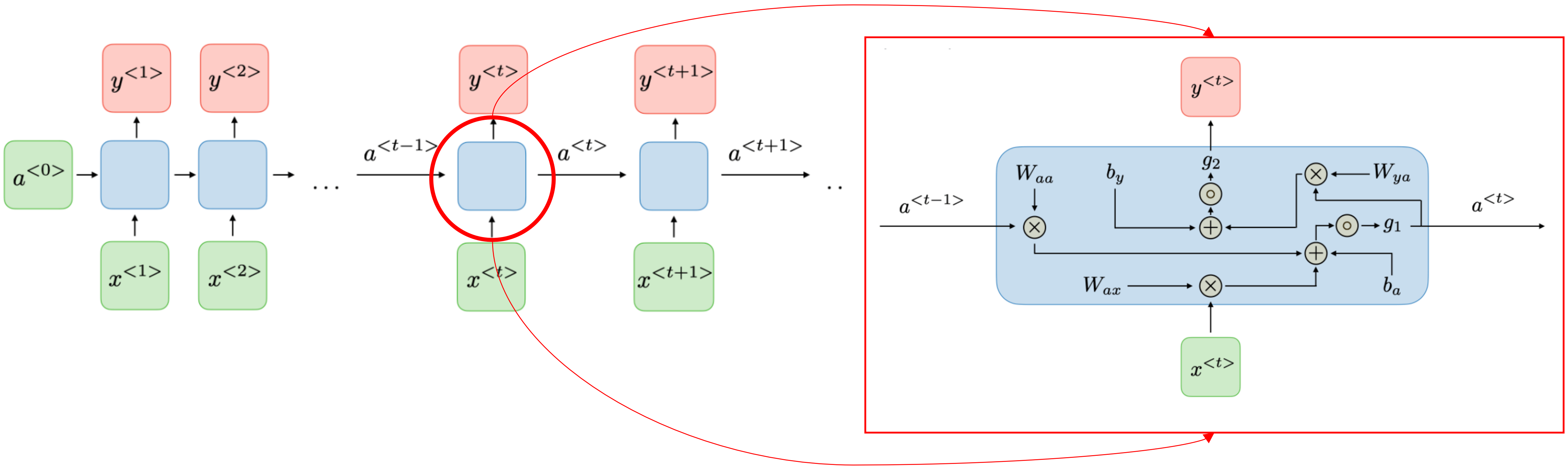
Neurons in 3 dimensions

- For images or data with spatial order
- Can stack up to >100 layers

http://cs231n.stanford.edu/

Recurrent Neural Networks

Traditional RNN

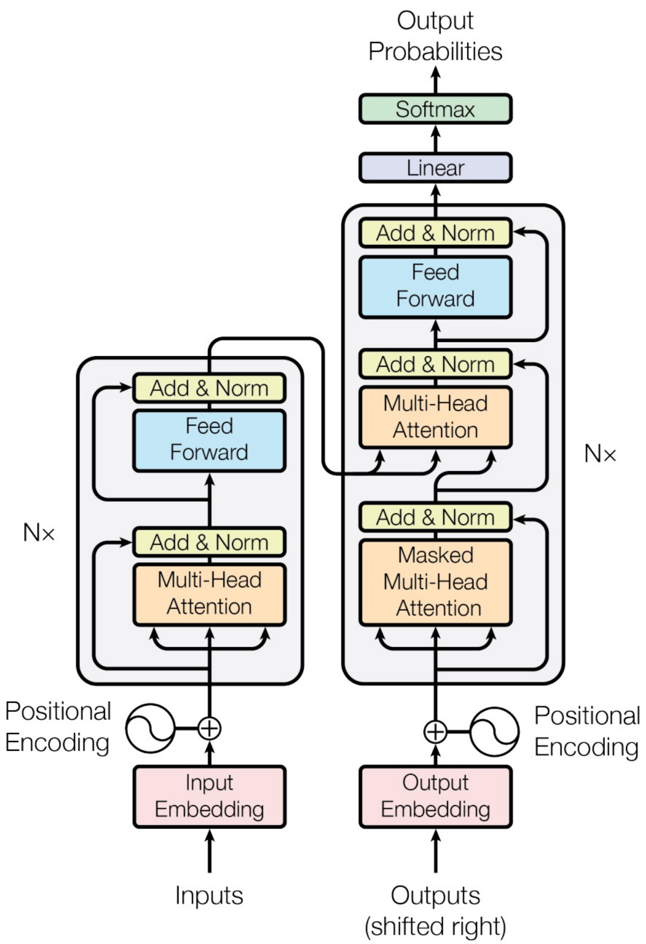


$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

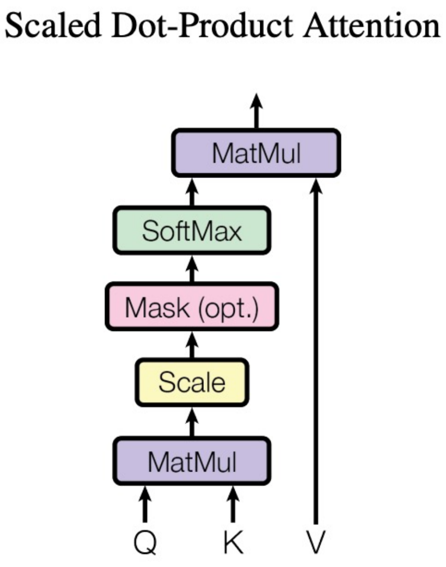
https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

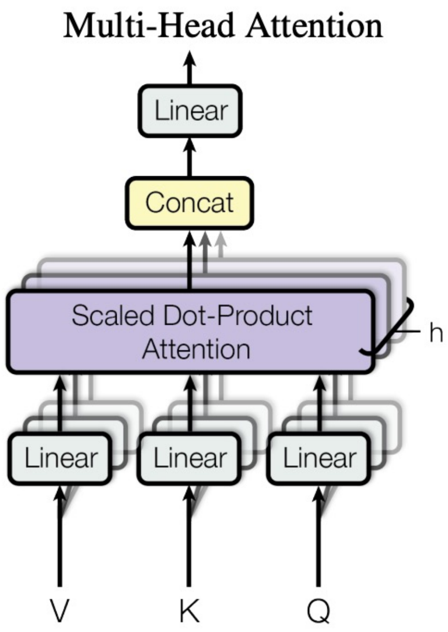
Transformers



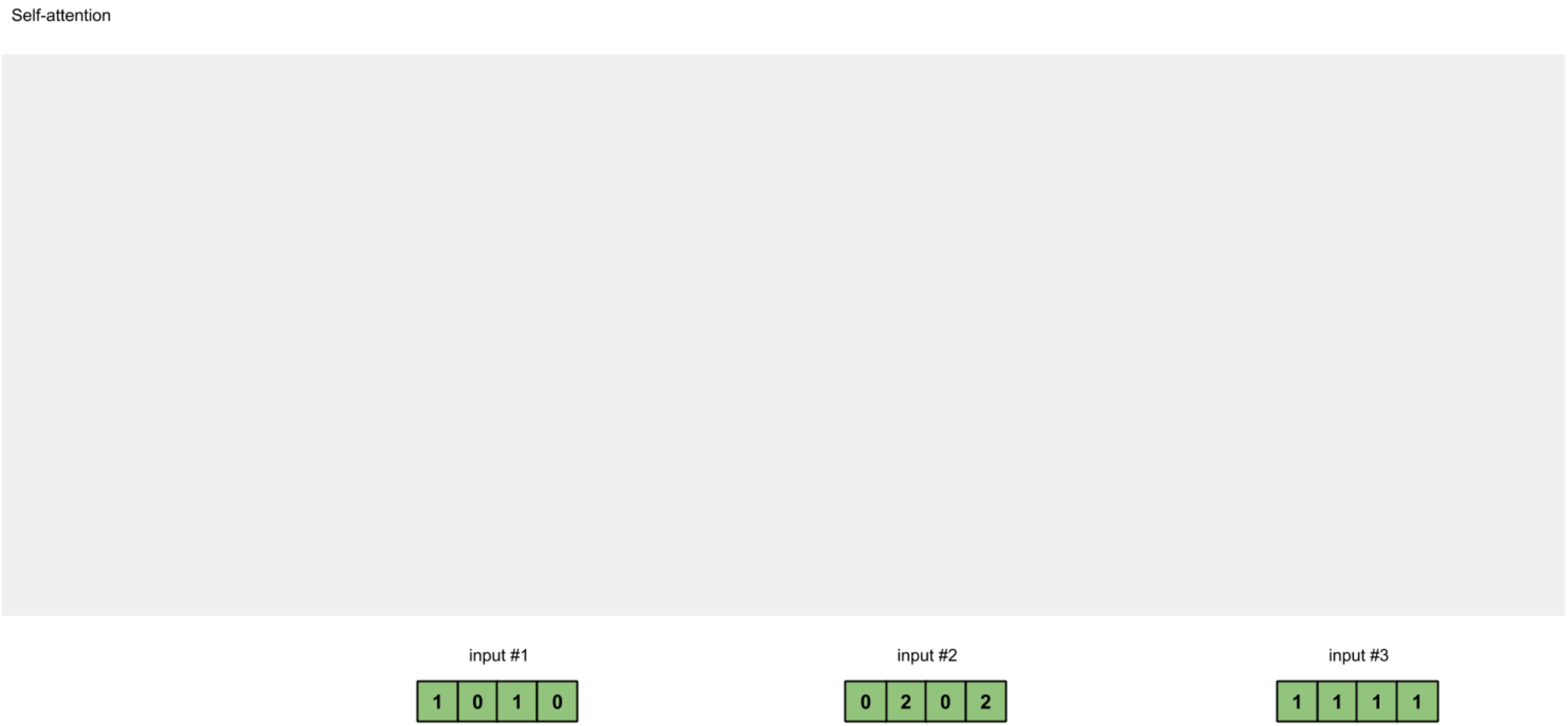
Encoder Decoder

Transformer: a new type of DNNs based on attention



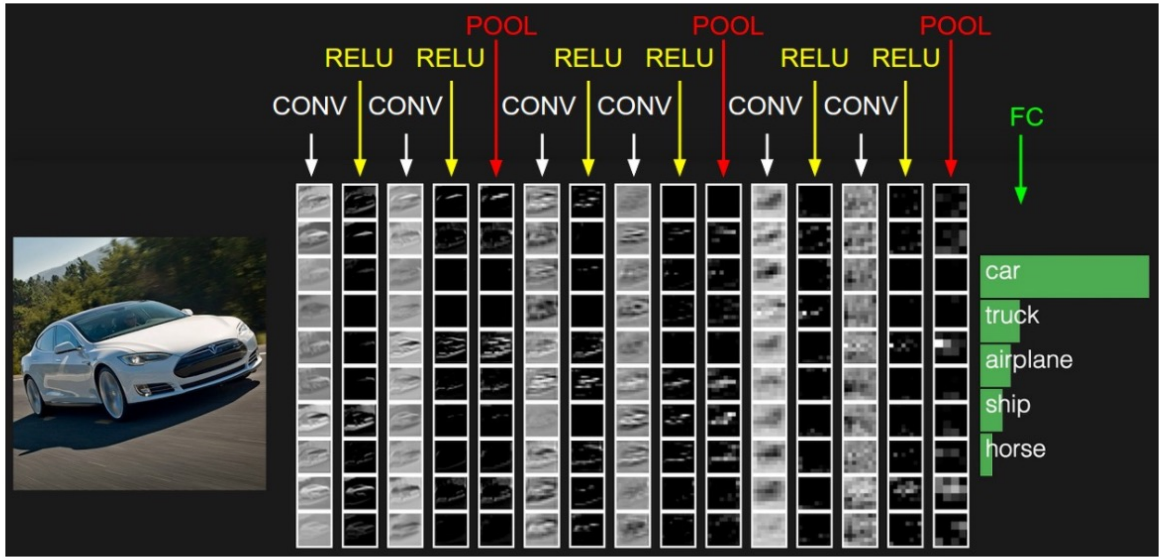


Self-Attention Explained



https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

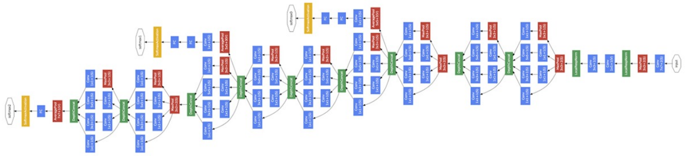
CNN Explained



A brief history of CNNs:

- **LeNet, 1990s**
- **AlexNet, 2012**
- **ZF Net, 2013**
- **GoogLeNet, 2014**
- **VGGNet, 2014**
- **ResNet, 2015**
- **Inception V4, 2016**
- **ResNeXt, 2017**
- **ViT, 2021**

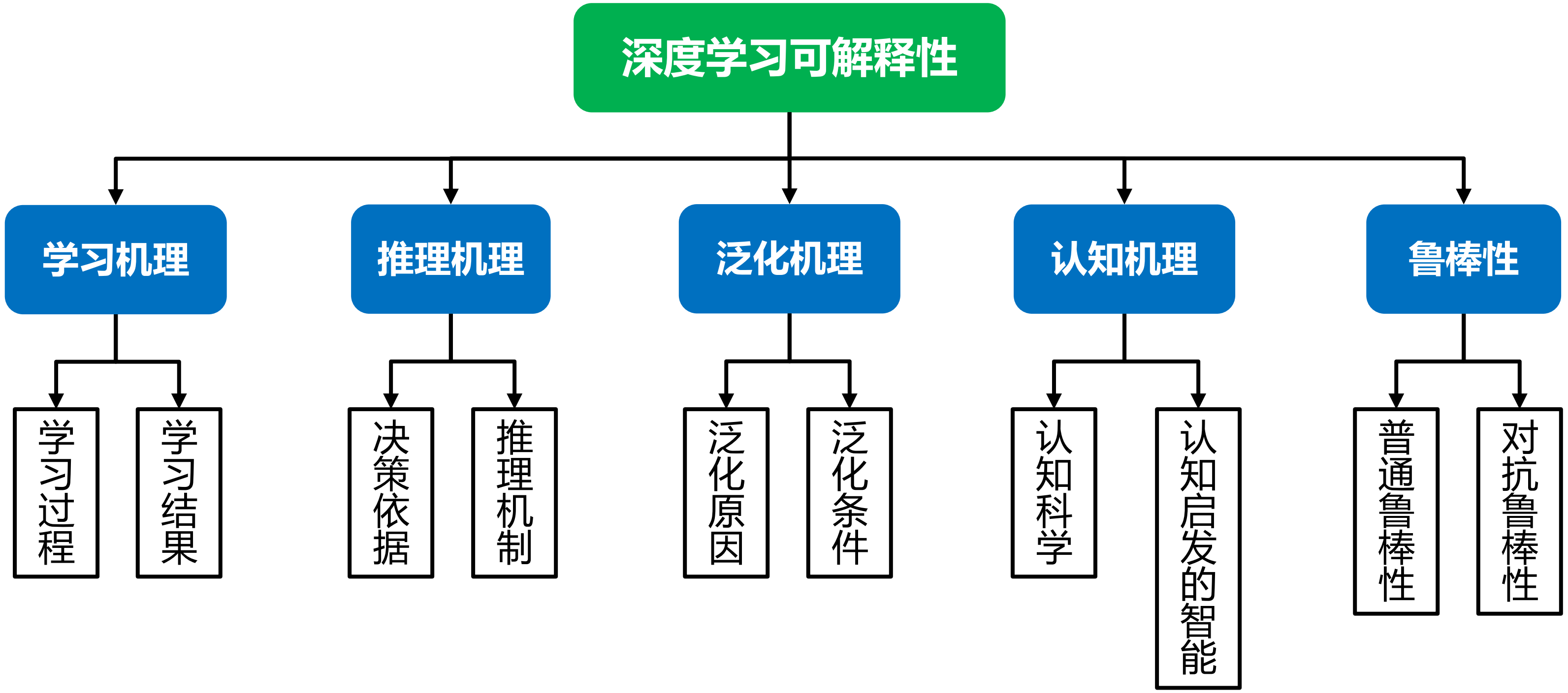
- Learns different levels of representations



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

http://cs231n.stanford.edu/

Explainable AI

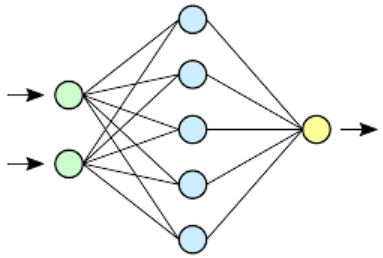


我们想要弄清楚下列问题：

- **DNN是怎么学习的、学到了什么、靠什么泛化、在什么情况下行又在什么情况下不行？**
- **深度学习是否是真正的智能，与人类智能比谁更高级，它的未来是什么？**
- **是否存在大一统的理论，不但能解释而且能提高？**

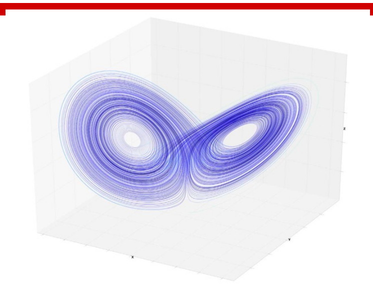
Methodological Principles

- ◆Visualization
- ◆Ablation
- ◆Contrast
- ◆Reverse



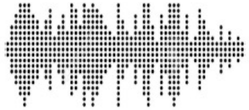
- Model
- Component
- Layer
- Operation
- Neuron

- Superclass
- Class
- Training/Test set
- Subset
- Sample



- Training
- Inference
- Transfer

How to Understand Machine Learning

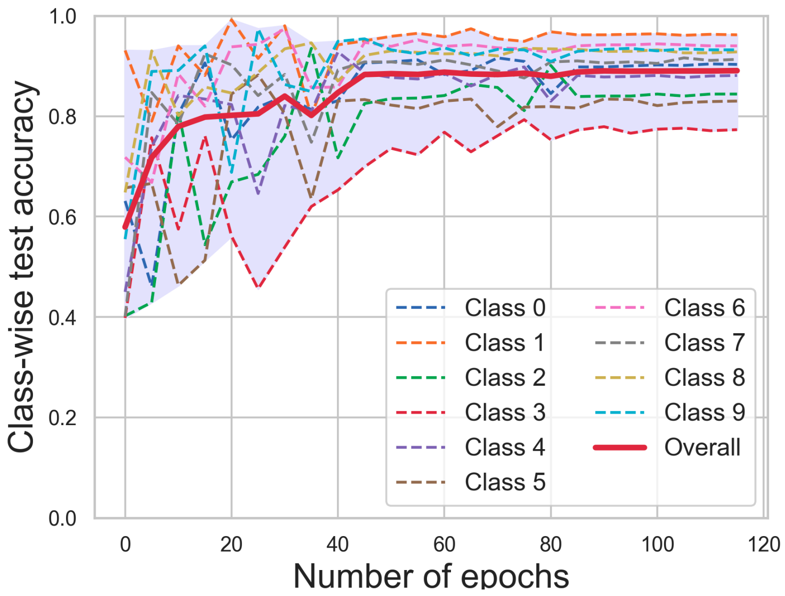
- ❖ 语音识别 f() = “天气不错”
- ❖ 人脸识别 f() = “小明”
- ❖ 语义分割 f() = 

Learning is the process of empirical risk minimization (ERM)

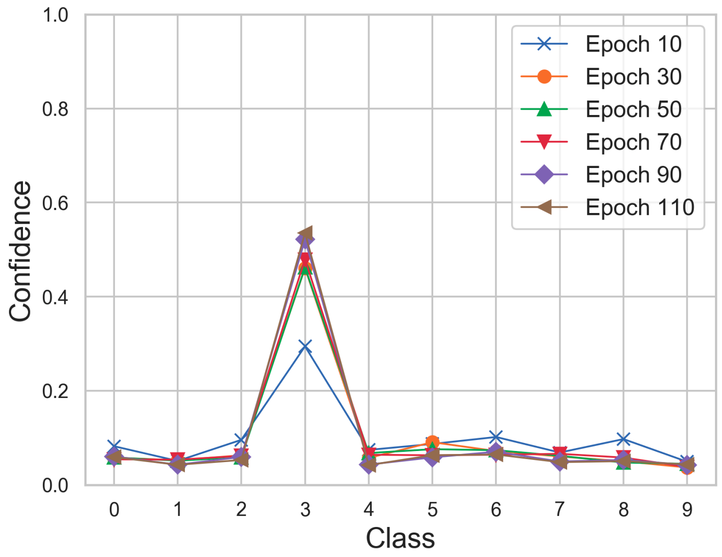
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\boldsymbol{x}_i), y_i)$$

Learning Mechanism

- Training/Test Error/Accuracy



- Prediction Confidence

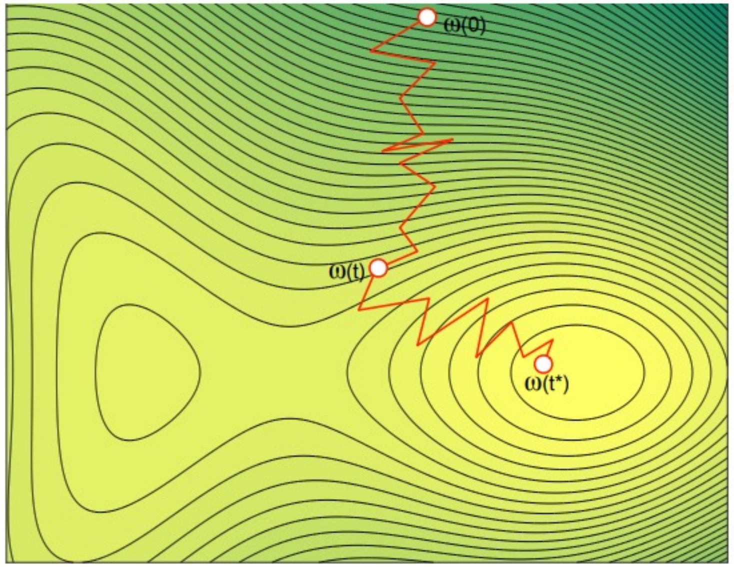


Explanation via observation: just plot!

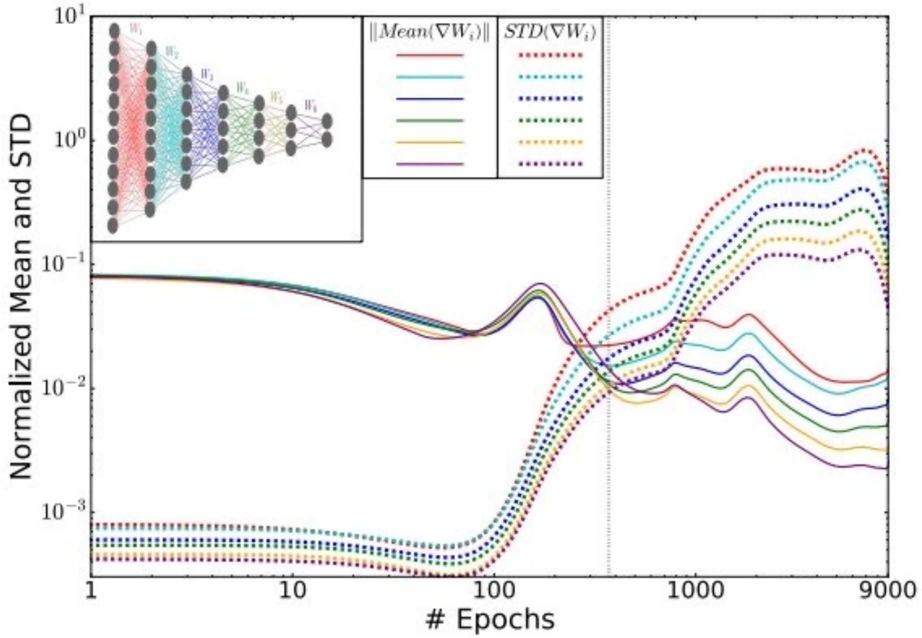
Wang et al. Symmetric Cross Entropy for Robust Learning with Noisy Labels, ICCV 2019.

Learning Mechanism

- **Parameter dynamics**



- **Gradient dynamics**

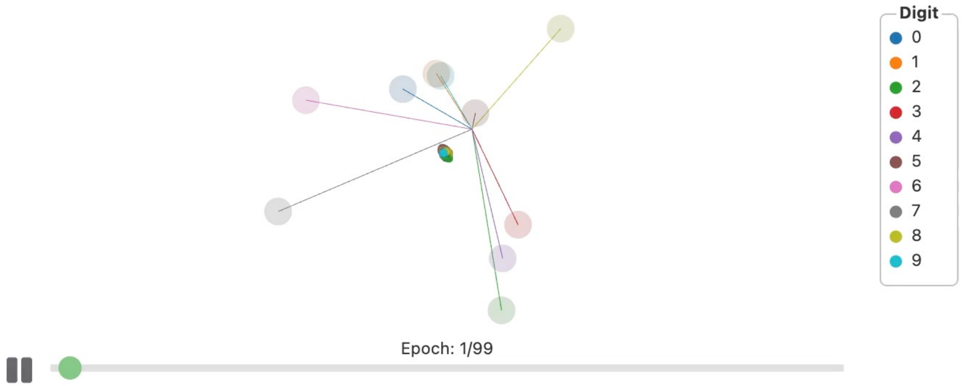


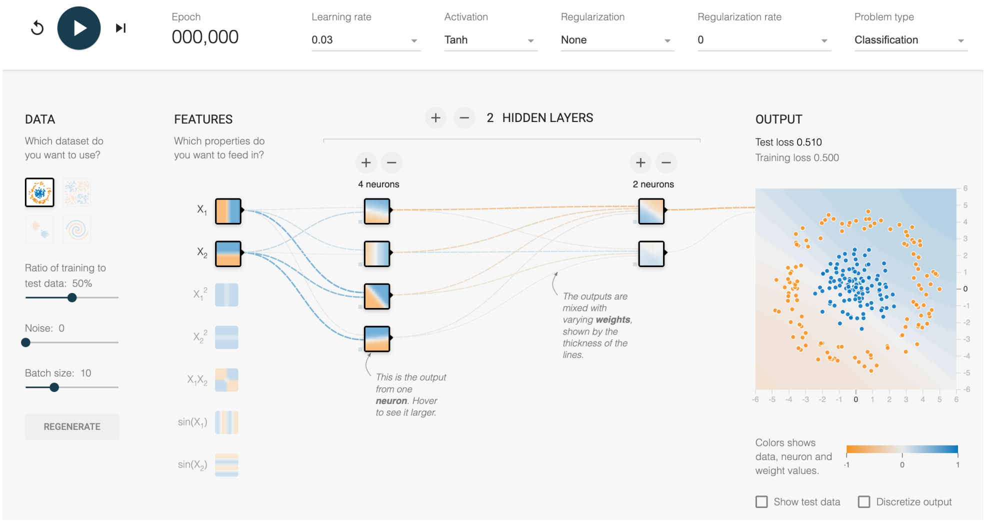
Explanation via dynamics and information

TRADI: Tracking deep neural network weight distributions, ECCV 2020; Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information[J]. arXiv:1703.00810, 2017.

Learning Mechanism

- Decision boundary, learning process visualization



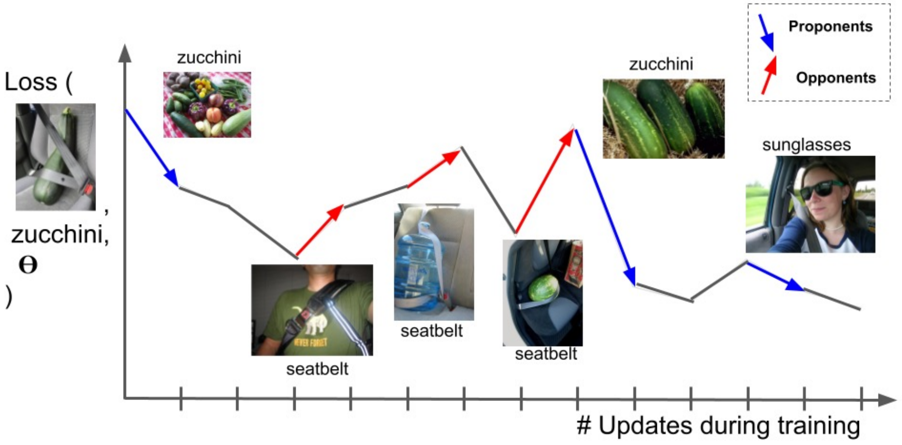


Explanation via dynamics and information

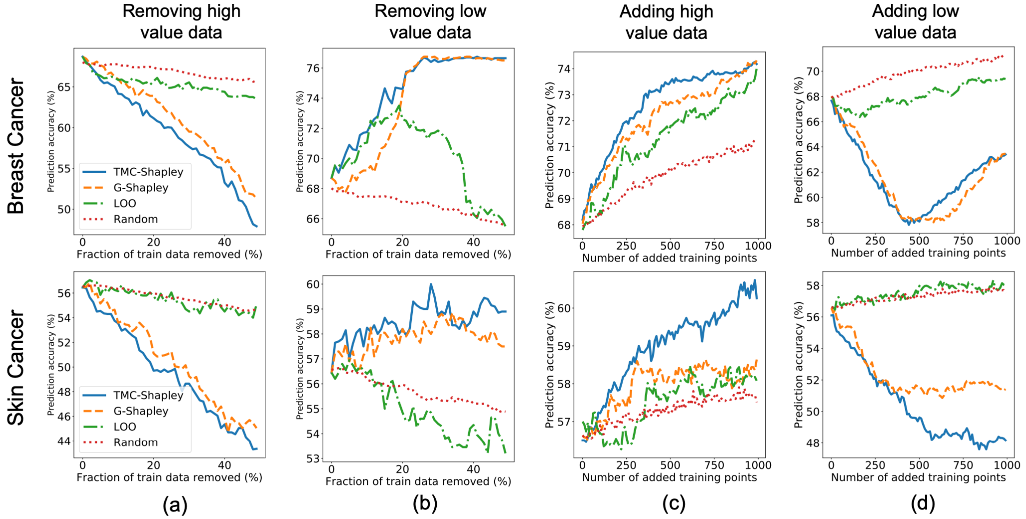
https://distill.pub/2020/grand-tour/ （March 16, 2020）; https://playground.tensorflow.org/

Learning Mechanism

- **Data influence/valuation: how a training sample impacts the learning outcome?**



Influence Function



Data Shapley

Understanding Black-box Predictions via Influence Functions, ICML, 2018;
Pruthi G, Liu F, Kale S, et al. Estimating training data influence by tracing gradient descent. NeurIPS, 2020.
Data shapley: Equitable valuation of data for machine learning, ICML, 2019.

Influence Function

- How model parameter would change if a sample z is removed from the training set?

目标： $\hat{\theta}_{-z} - \hat{\theta}$ $\quad\quad$ $\hat{\theta}_{-z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \sum_{z_i \neq z} L(z_i, \theta)$

- How model parameter would change if z is upweighted by a small constant ϵ?

$$\mathcal{I}_{\text{up,params}}(z) \stackrel{\text{def}}{=} \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z,\hat{\theta}) \quad H_{\hat{\theta}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i,\hat{\theta})$$

Cook, R. D. and Weisberg, S. Residuals and influence in regression. New York: Chapman and Hall, 1982

- Removing sample z is equivalent to upweighting it by $\epsilon = -\frac{1}{n}$

$O(np^2 + p^3)$

所以： $\hat{\theta}_{-z} - \hat{\theta} \approx -\frac{1}{n}\mathcal{I}_{\text{up,params}}(z)$

complexity=O(#samples*$\#\theta^2 + \#\theta^3$)

Training Data Influence

- How model loss on z′ would change if update on a sample z?

$$\texttt{TracInIdeal}(z, z') = \sum_{t:\ z_t = z} \ell(w_t, z') - \ell(w_{t+1}, z')$$

- First-order approximation of the above (assuming one step update is small)?

$$\ell(w_{t+1}, z') = \ell(w_t, z') + \nabla \ell(w_t, z') \cdot (w_{t+1} - w_t) + O(\|w_{t+1} - w_t\|^2)$$

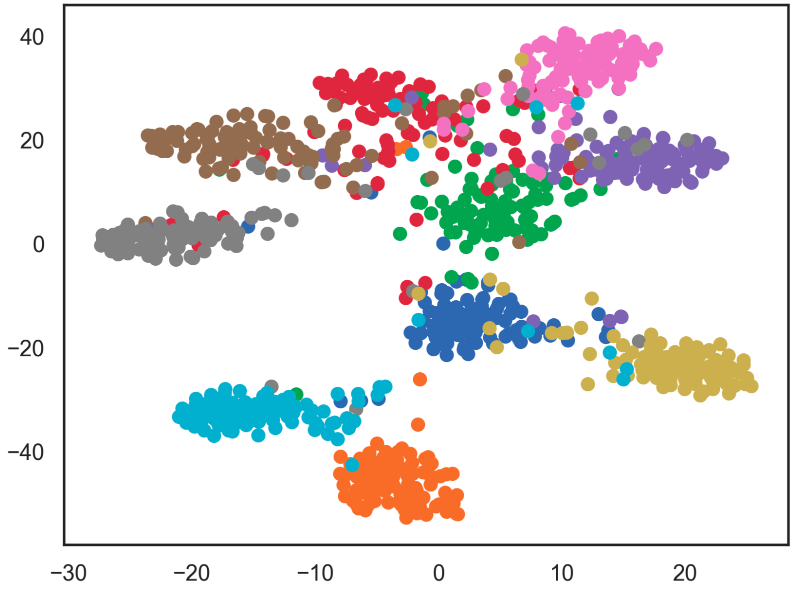
$$w_{t+1} - w_t = -\eta_t \nabla \ell(w_t, z_t)$$

- Checkpoints store the interim updates

所以：$$\texttt{TracInCP}(z, z') = \sum_{i=1}^{k} \eta_i \nabla \ell(w_{t_i}, z) \cdot \nabla \ell(w_{t_i}, z')$$

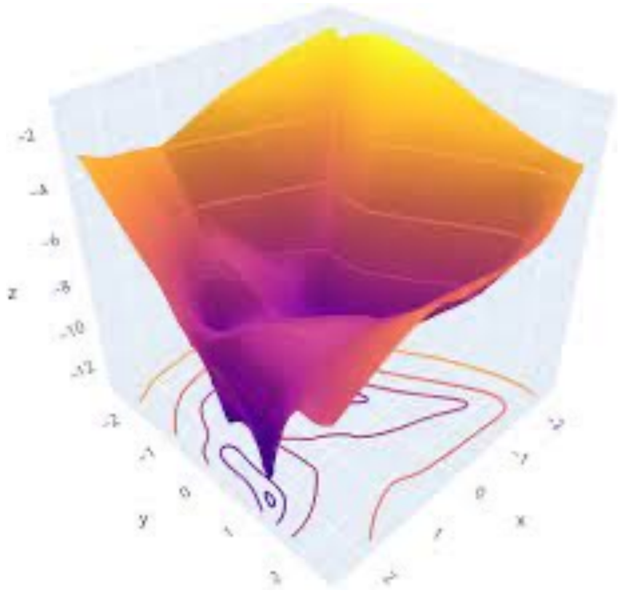
Understanding the Learned Model

◻ Deep features



t-SNE plot

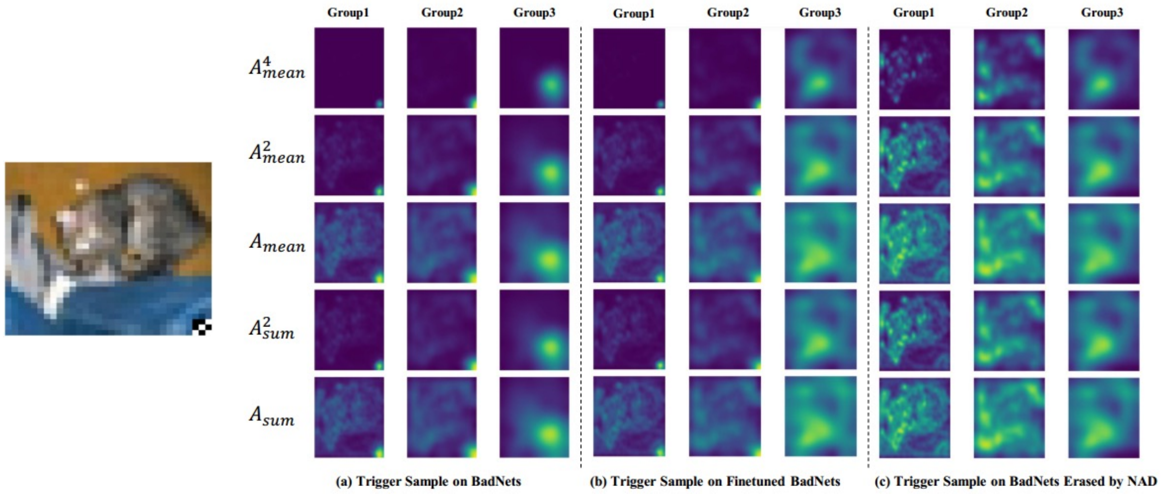
◻ Loss Landscape



Maaten et al. Visualizing data using t-SNE. JMLR, 2008.
https://distill.pub/2016/misread-tsne/?_ga=2.135835192.888864733.1531353600-1779571267.1531353600

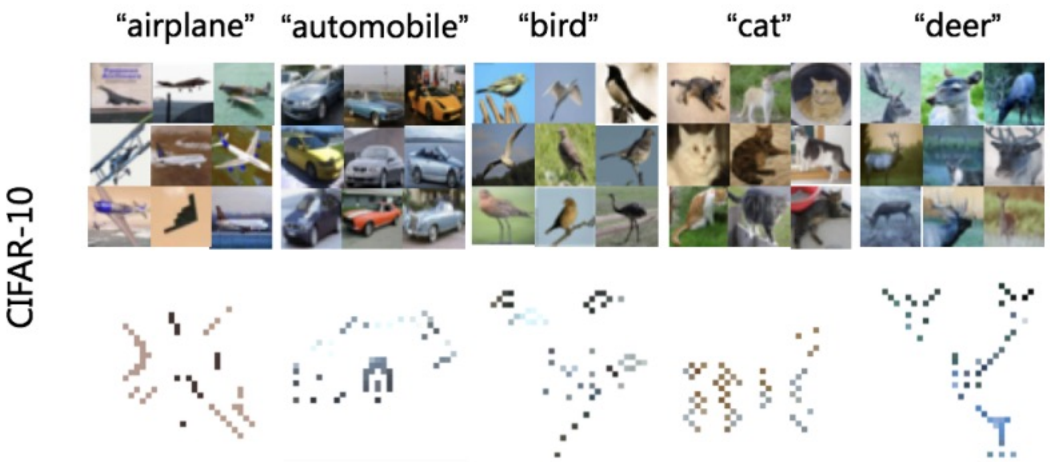
Understanding the Learned Model

Intermediate Layer Activation Map



Activation/Attention Map

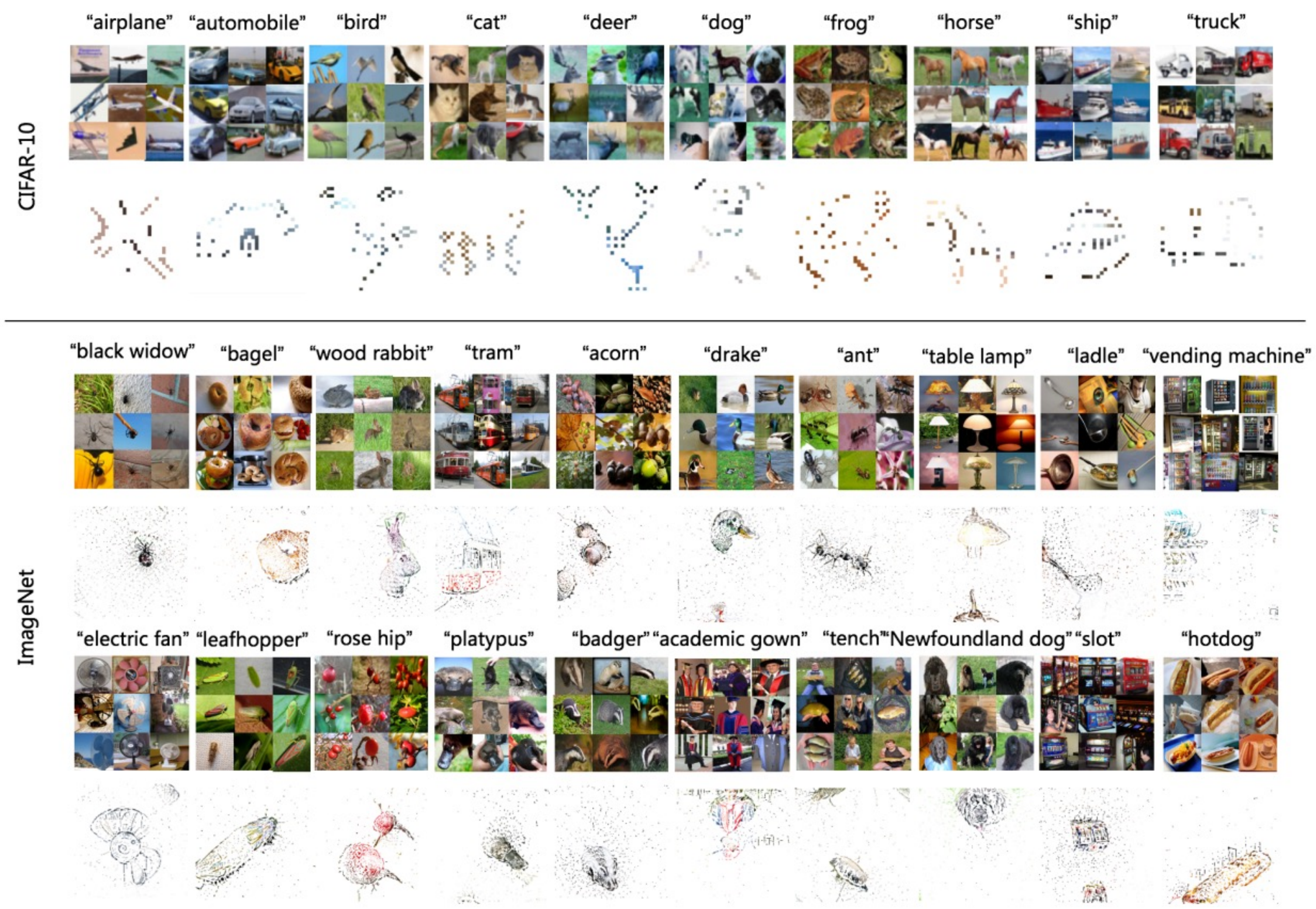
Class-wise Patterns



One predictive pattern for each class

Li et al. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Network, ICLR 2021; Zhao et al. What do deep nets learn? class-wise patterns revealed in the input space. *arXiv:2101.06898* (2021).

What do deep nets learn?



Goal: understanding knowledge learned by a model of a particular class.

Method: Extract one single pattern for one class, then what this pattern would be?

Other considerations: we need to do this in **pixel space**, as they are more interpretable

Zhao, Shihao, et al. "What do deep nets learn? class-wise patterns revealed in the input space." *arXiv:2101.06898* (2021).

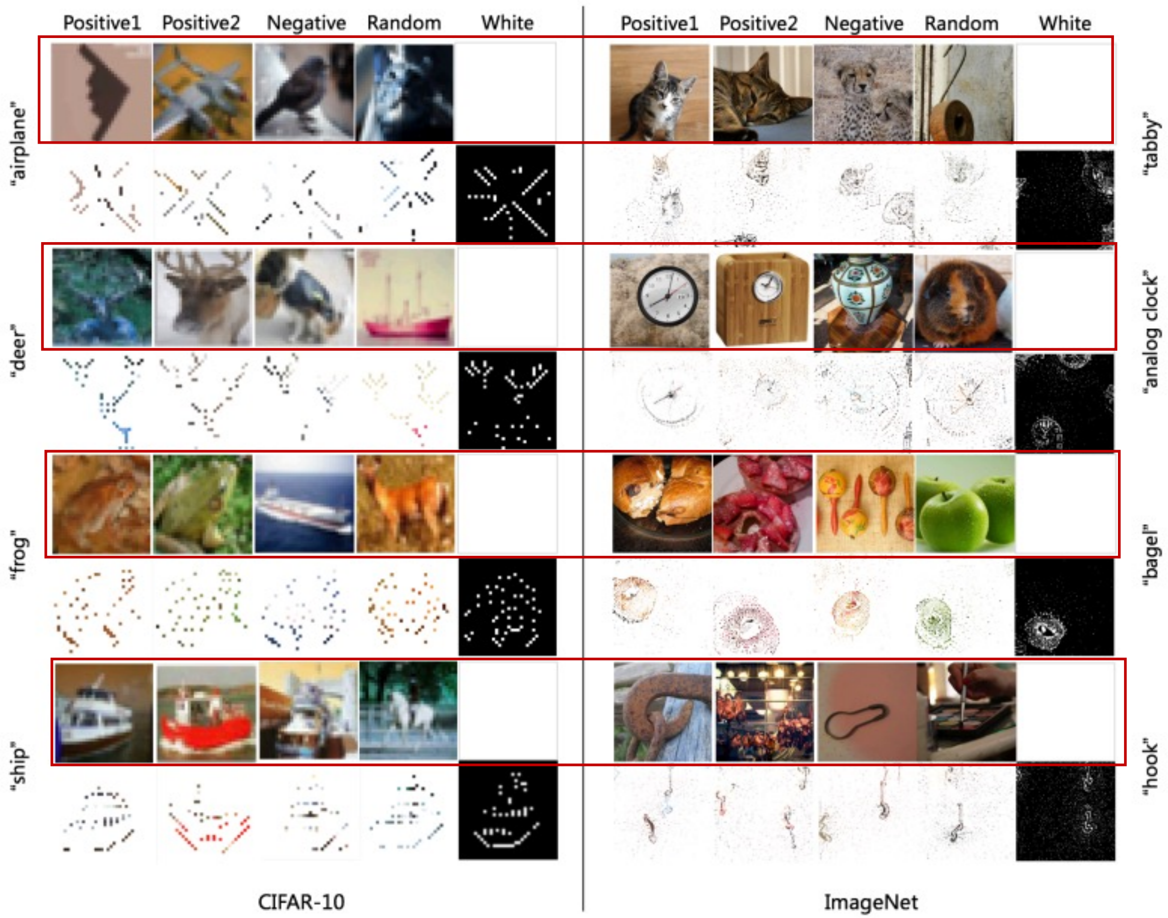
How to Find the Class-wise Pattern

$$\mathcal{L} = -\log f_y(\tilde{\boldsymbol{x}}) + \alpha \frac{1}{n} \|\boldsymbol{m}\|_1$$

$$\tilde{\boldsymbol{x}} = \boldsymbol{m} * \boldsymbol{x}_c + (1 - \boldsymbol{m}) * \boldsymbol{x}_n$$

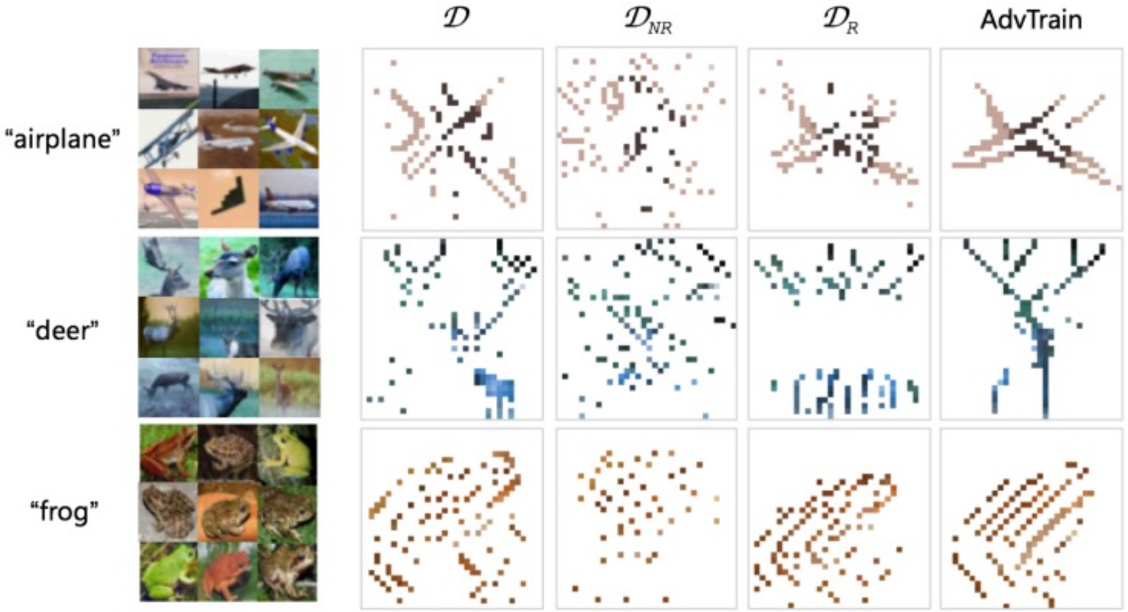
$$\boldsymbol{x}_n \in \mathcal{D}_n \subset \mathcal{D}_{test}$$

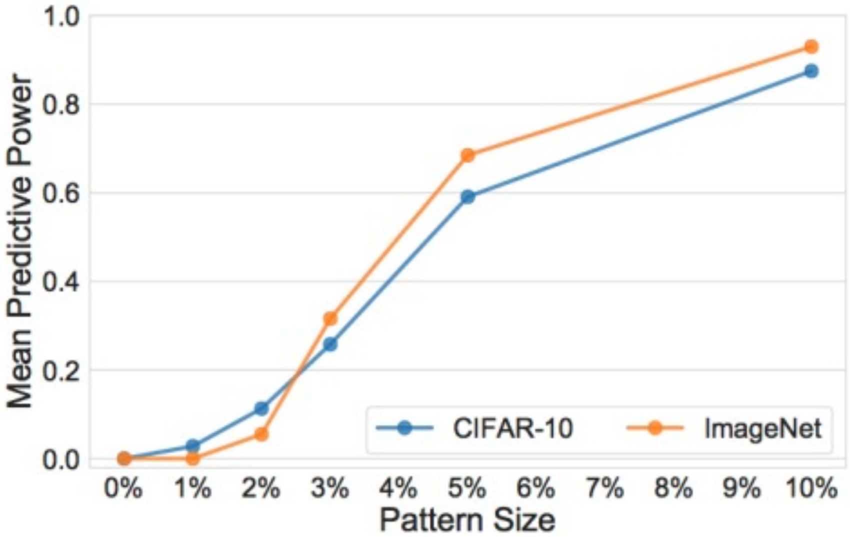
$\boldsymbol{x}_c$: a canvas image



Patterns extracted on different canvases (red rectangles)

Class-wise Patterns Revealed





Patterns extracted on original, non-robust, robust CIFAR-10and patterns of adversarially trained models

Predictive power of different sizes of patterns

Inference Mechanism

☐ Guided Backpropagation

(a) Original Image (b) Guided Backprop 'Cat'

☐ Class Activation Map (Grad-CAM)

Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV 2017.
Springenberg et al. Striving for Simplicity: The All Convolutional Net, ICLR 2015.

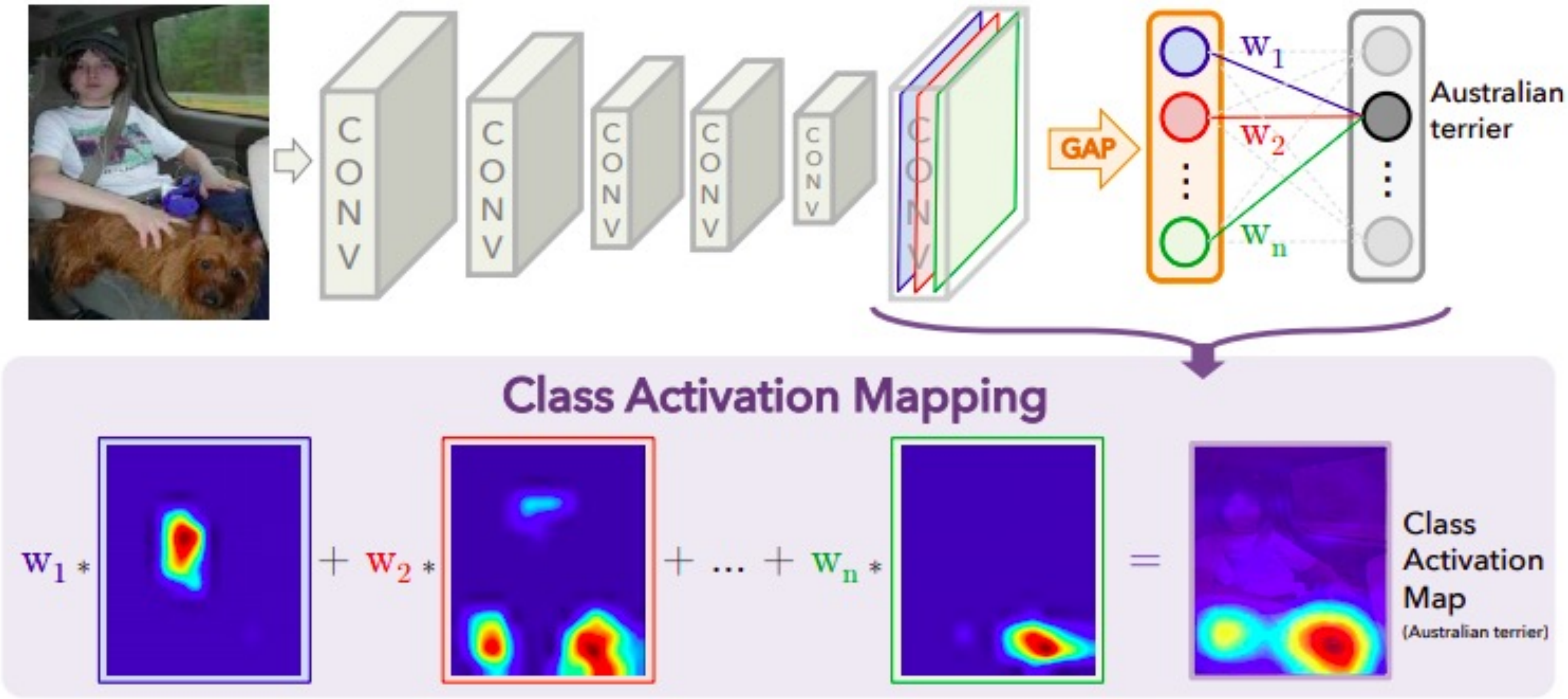
Guided Backpropagation

ReLU forward pass	$h^{l+1} = \max\{0, h^l\}$	Forward pass	$h^l \begin{pmatrix} 1 & -1 & 5 \\ 2 & -5 & -7 \\ -3 & 2 & 4 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 5 \\ 2 & 0 & 0 \\ 0 & 2 & 4 \end{pmatrix} h^{l+1}$
ReLU backward pass	$\frac{\partial L}{\partial h^l} = [[h^l > 0]] \frac{\partial L}{\partial h^{l+1}}$	Backward pass: backpropagation	$\begin{pmatrix} -2 & 0 & -1 \\ 6 & 0 & 0 \\ 0 & -1 & 3 \end{pmatrix} \leftarrow \begin{pmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{pmatrix} \frac{\partial L}{\partial h^{l+1}}$
Deconvolution for ReLU	$\frac{\partial L}{\partial h^l} = [[h^{l+1} > 0]] \frac{\partial L}{\partial h^{l+1}}$	Backward pass: “deconvnet”	$\begin{pmatrix} 0 & 3 & 0 \\ 6 & 0 & 1 \\ 2 & 0 & 3 \end{pmatrix} \leftarrow \begin{pmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{pmatrix} \frac{\partial L}{\partial h^{l+1}}$
Guided Backpropagation	$\frac{\partial L}{\partial h^l} = [[(h^l > 0)\&\&(h^{l+1} > 0)]] \frac{\partial L}{\partial h^{l+1}}$	Backward pass: *guided backpropagation*	$\begin{pmatrix} 0 & 0 & 0 \\ 6 & 0 & 0 \\ 0 & 0 & 3 \end{pmatrix} \leftarrow \begin{pmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{pmatrix} \frac{\partial L}{\partial h^{l+1}}$

Springenberg et al. Striving for Simplicity: The All Convolutional Net, ICLR 2015.
https://medium.com/@chinesh4/generalized-way-of-interpreting-cnns-a7d1b0178709

Class Activation Mapping (CAM)

GAP: Global Average Pooling



Zhou et al. Learning Deep Features for Discriminative Localization. CVPR, 2016.
https://medium.com/@chinesh4/generalized-way-of-interpreting-cnns-a7d1b0178709

Grad-CAM

Grad-CAM is a generalization of CAM

Compute **neuron importance**:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

y^c : logits of class c (before softmax)

A^k : k-th channel activation map

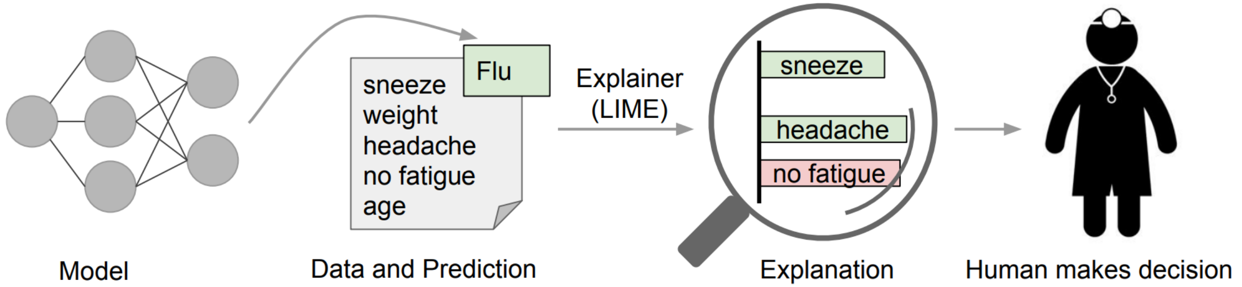
Weighted combination of activation map, then **interpolation**:

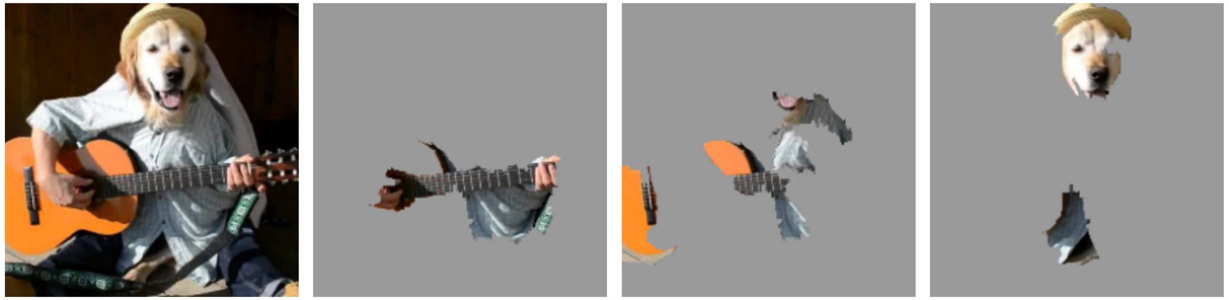
$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In CVPR, 2016; https://medium.com/@chinesh4/generalized-way-of-interpreting-cnns-a7d1b0178709

LIME

Local Interpretable Model-agnostic Explanations (LIME)





(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

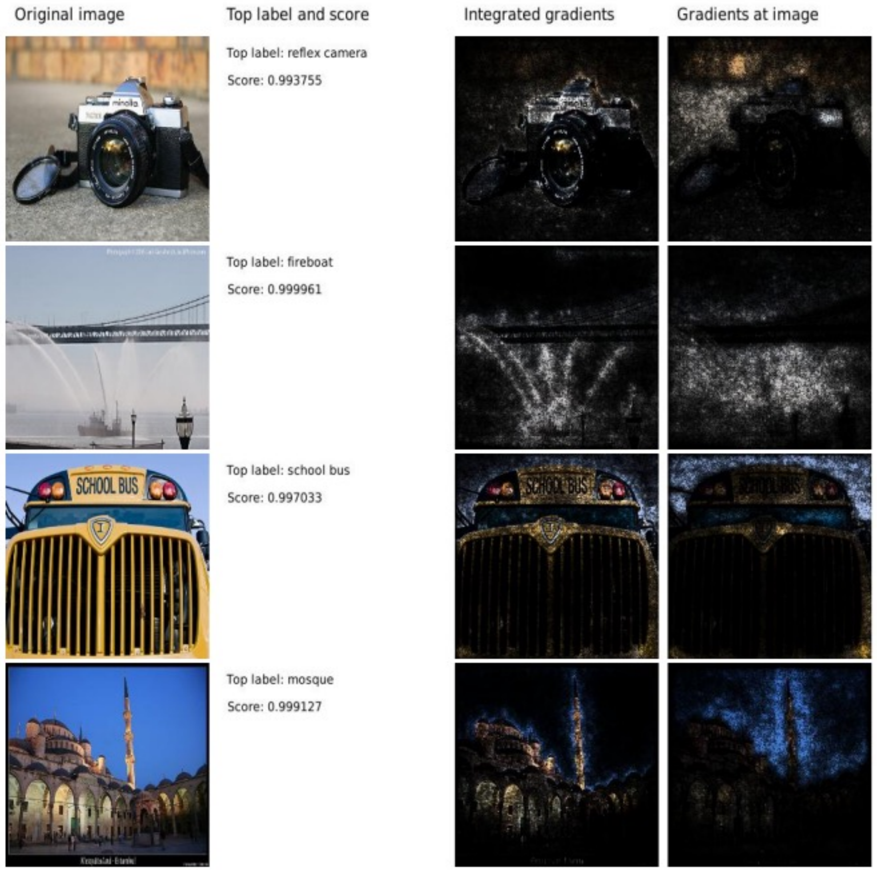
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

π_x : local neighborhood of x
z : sampled neighbor points
g : explainer e.g a linear model
z' : a binary vector for interpretable representation(e.g. patch)

Ribeiro et al. “Why should i trust you?” Explaining the predictions of any classifier. “ SIGKDD, 2016.
https://github.com/marcotcr/lime

Integrated Gradients

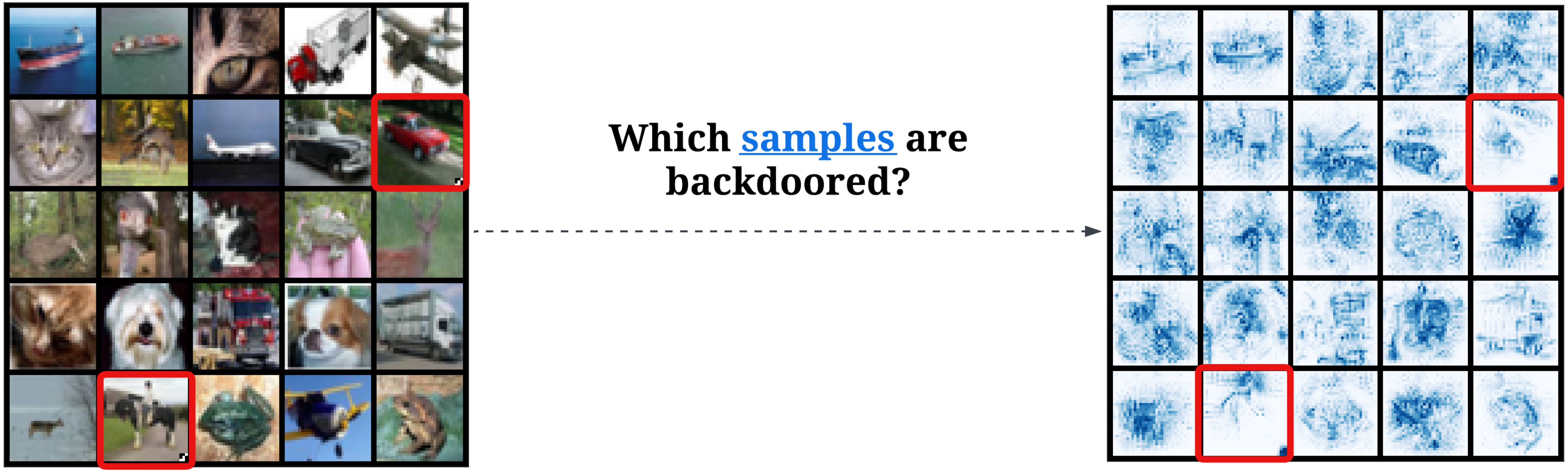


$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

- **There is a path**: $x_i \rightarrow x'_i$
- **Traverse the path using** α
- **Integrate the gradients along the way**

Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks, ICML, 2017.
https://github.com/TianhongDai/integrated-gradient-pytorch

Cognitive Distillation



- Mask extract by cognitive distillation

Useful and non-useful features

- **Useful features**:

➢ highly correlated with the true label in expectation, so
  - If removed, prediction change
  - Backdoor trigger is a useful feature

- **Non-useful features**:

➢ not correlated with prediction
  - If removed, prediction does not change

Ilyas, Andrew, et al. "Adversarial examples are not bugs, they are features.” NeurIPS 2019

Cognitive Distillation

Objective: distill the minimal essence of useful features

$$\arg\min_{\boldsymbol{m}} \| f_\theta(\boldsymbol{x}) - f_\theta(\boldsymbol{x}_{cp}) \|_1 + \alpha \|\boldsymbol{m}\|_1 + \beta TV(\boldsymbol{m})$$

$$\boldsymbol{x}_{cp} = \boldsymbol{x} \odot \boldsymbol{m} + (1 - \boldsymbol{m}) \odot \delta$$

Cognitive Distillation

Model

Total Variation Loss

$$\arg\min_{\boldsymbol{m}} \boxed{\|f_\theta(\boldsymbol{x}) - f_\theta(\boldsymbol{x}_{cp})\|_1} + \alpha\|\boldsymbol{m}\|_1 + \boxed{\beta TV(\boldsymbol{m})}$$

$$\boldsymbol{x}_{cp} = \boldsymbol{x} \odot \boldsymbol{m} + (1 - \boldsymbol{m}) \odot \delta,$$

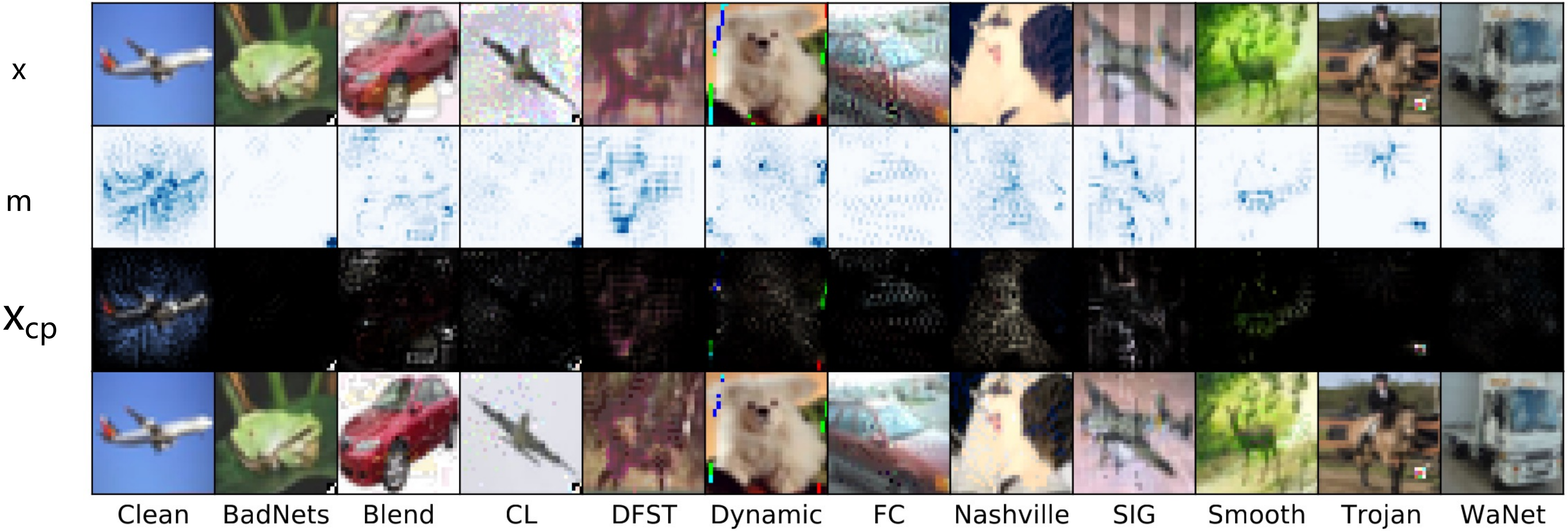
Cognitive Pattern

Original image

Mask

Random noise vector

Distilled patterns on backdoored samples



How to Verify Cognitive Patterns are Essential

Construct simplified backdoor patterns:

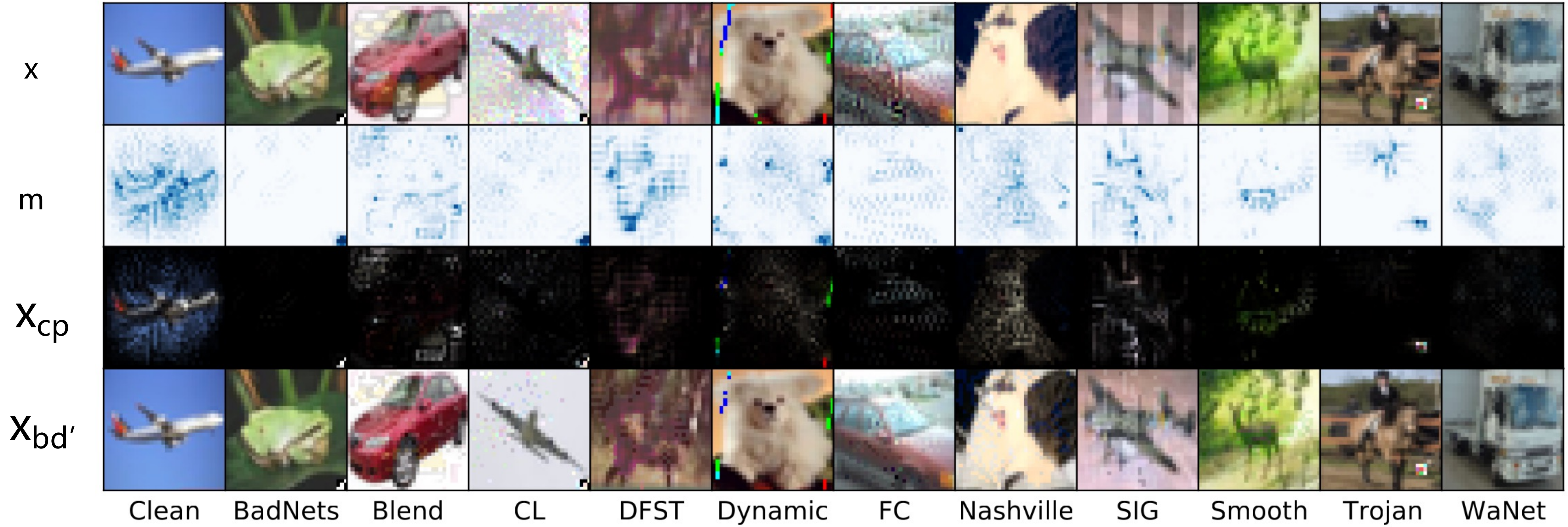
Backdoored image

$$\boldsymbol{x}'_{bd} = \boldsymbol{m} \odot \boldsymbol{x}_{bd} + (1 - \boldsymbol{m}) \odot \boldsymbol{x},$$

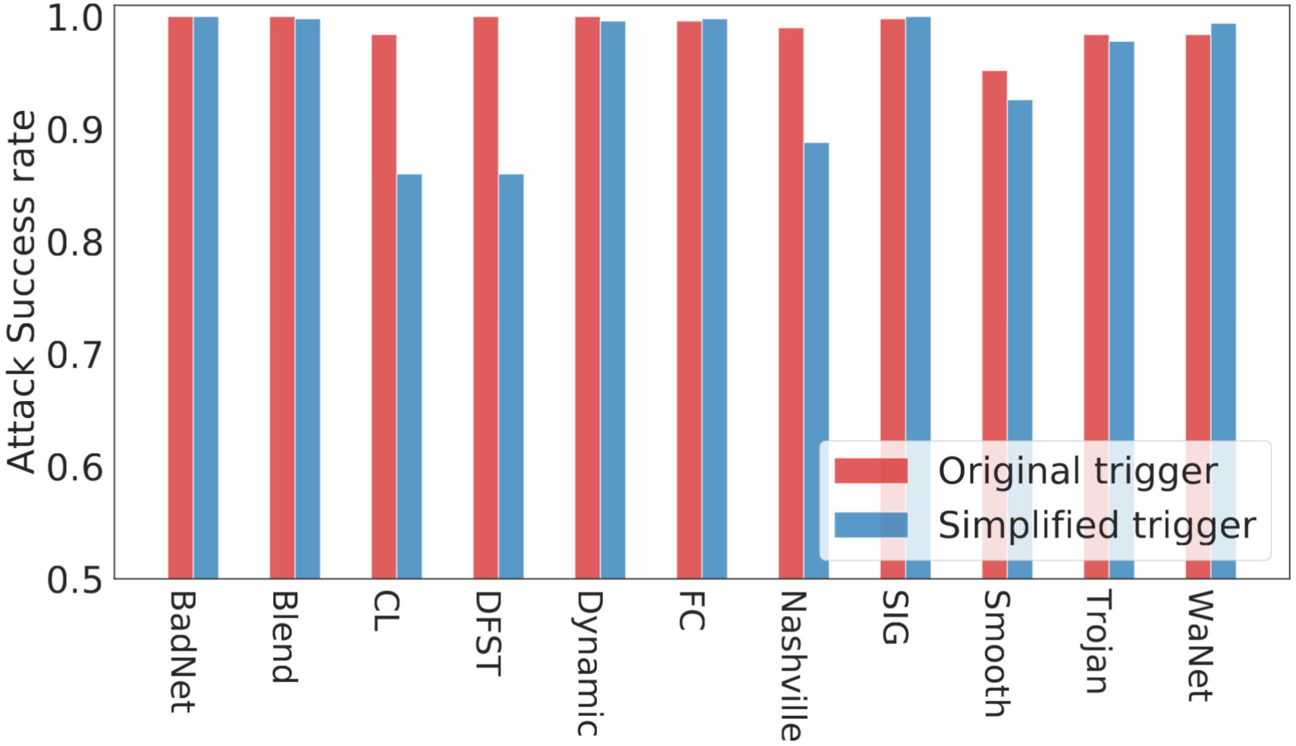
Binarized mask {0,1}

Original image

Backdoor Patterns Can Be Made Simpler

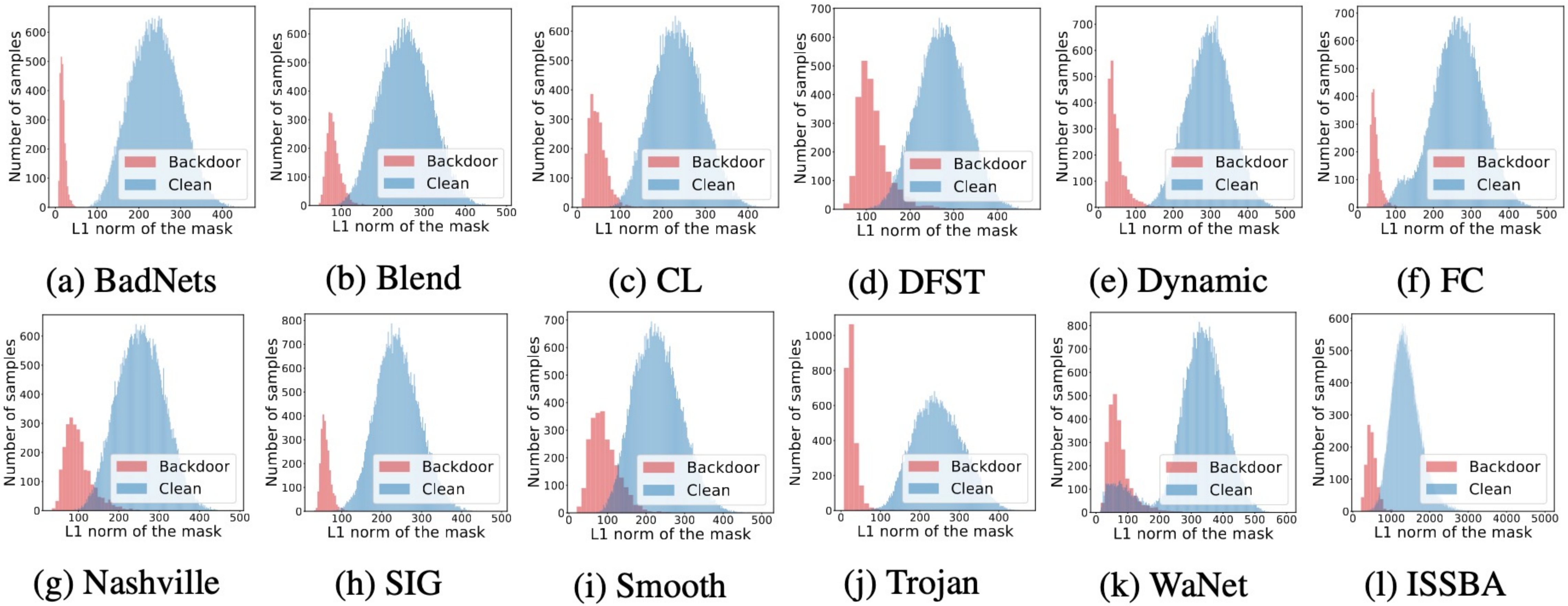


Backdoor Patterns Can Be Made Simpler



Simplified backdoor patterns also work!

L1 Norm Distribution of the Distilled Mask



(a) BadNets (b) Blend (c) CL (d) DFST (e) Dynamic (f) FC

(g) Nashville (h) SIG (i) Smooth (j) Trojan (k) WaNet (l) ISSBA

Detect Backdoor Samples

- **Attacks**: 12 backdoor attacks
- **Models** : ResNet-18, Pre-Activation ResNet-101, MobileNet v2, VGG-16, Inception, EfficientNet-b0
- **Datasets**: CIFAR-10 / GTSRB / ImageNet subset
- **Evaluation metric**: area under the ROC curve (AUROC)
- **Detection baselines**:
  - Anti-Backdoor Learning (ABL) [2]
  - Activation Clustering (AC) [3]
  - Frequency [4]
  - STRIP [5]
  - Spectral Signatures [6]
- CD-L (logits layer) and CD-F (last activation layer)

Superb Detection Performance

Table 1: The detection AUROC (%) of our CD method and the baselines against 12 backdoor attacks (poisoning rate 5%) on the *training/test* set. The results are averaged across the 6 models (VGG-16, RN-18, PARN-101, MobileV2, GoogLeNet, and EfficientNet-b0). The best results are in **bold**.

Dataset	Attack	ABL	AC	Frequency	STRIP	SS	CD-L	CD-F
CIFAR10	BadNets	85.64/-	77.57/74.63	92.32/91.59	**97.89/97.66**	62.89/45.50	94.03/94.72	88.89/89.88
	Blend	88.17/-	76.23/65.93	80.67/79.40	84.55/83.02	51.63/40.52	**93.47/93.44**	92.30/92.41
	CL	90.86/-	70.06/25.68	**98.85**/91.59	97.27/**96.04**	40.78/39.02	98.75/85.31	93.48/80.31
	DFST	**89.10**/-	80.45/86.97	87.62/87.34	58.08/58.51	56.34/40.69	88.96/**89.80**	82.54/82.68
	Dynamic	87.97/-	77.83/77.07	97.82/97.58	91.49/89.75	66.49/50.91	**97.97/97.85**	94.89/94.76
	FC	86.61/-	83.99/88.74	98.65/98.11	79.84/76.97	63.62/64.62	**99.17/98.22**	94.46/95.12
	SIG	**97.42**/-	84.40/56.91	62.95/56.46	81.68/57.44	58.90/52.70	96.91/90.90	96.09/**93.17**
	Smooth	79.53/-	82.11/76.48	51.32/47.84	58.52/55.81	70.24/51.14	**91.09/89.03**	82.05/81.91
	Nashville	76.12/-	89.26/76.11	70.53/67.71	51.62/48.30	80.48/60.62	**98.10/97.34**	95.28/94.26
	Trojan	85.96/-	69.59/71.58	93.82/93.36	91.85/92.14	59.18/45.04	**96.91/96.72**	91.16/91.88
	WaNet	56.66/-	70.96/69.86	**96.31/96.65**	84.98/84.64	71.59/57.27	95.69/96.08	86.60/88.43
GTSRB	BadNets	67.78/-	98.21/72.79	-	57.26/59.59	69.97/72.86	99.28/99.14	**99.59/99.66**
ImageNet	BadNets	83.40/-	95.75/**100.00**	-	96.05/95.84	99.73/9.20	**100.00/100.00**	**100.00/100.00**
	ISSBA	96.99/-	**100.00**/80.29	-	70.37/68.73	42.22/56.31	**100.00/99.99**	99.97/99.89
Average	-	83.61/-	82.60/73.21	84.62/82.51	78.61/75.96	63.83/49.58	**96.45/94.90**	92.66/91.74

Discover Biases in Facial Recognition Models

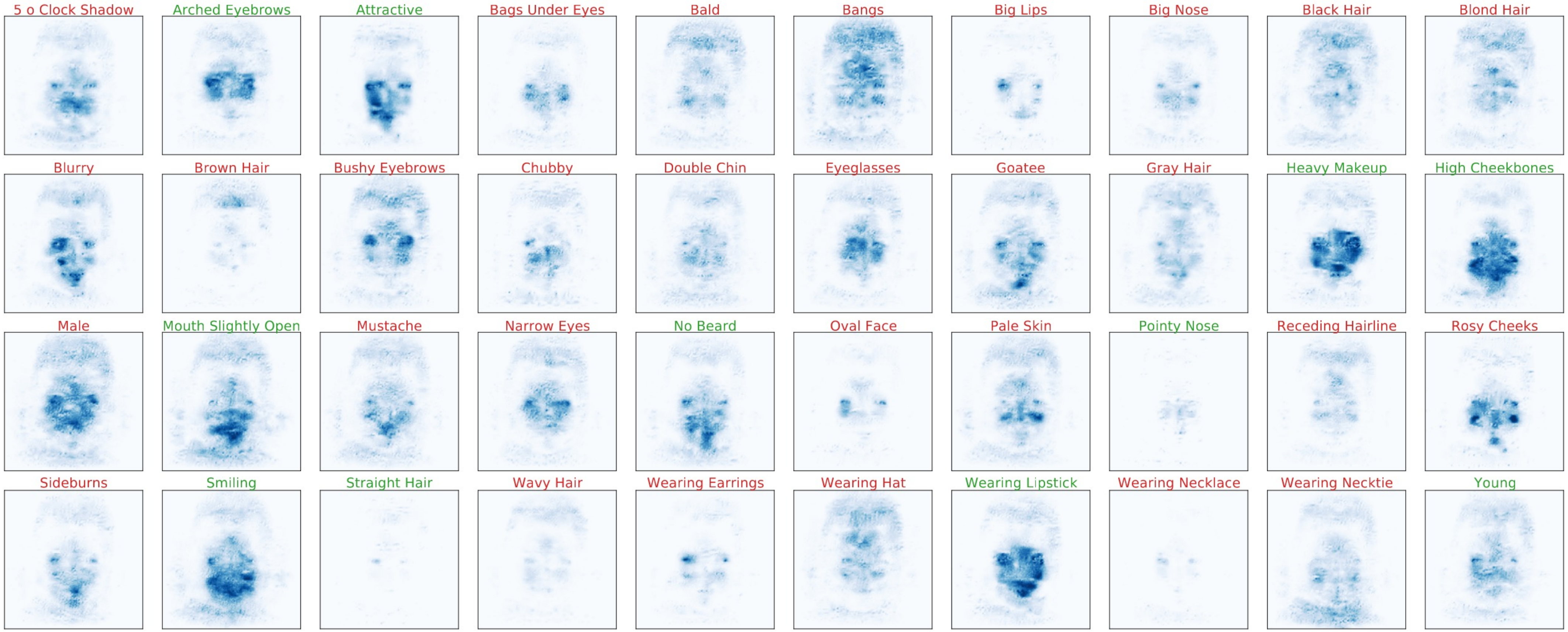
CelebA dataset:

- 40 binary facial attributes (gender, bald, and hair color)
- Known bias between *gender* and *blond hair*

- Apply **CD** in the same way as backdoor detection
  - Select subset of samples with low L1 norm
  - Examine attributes of the subset
  - Calculate distribution shift between subset and the full dataset

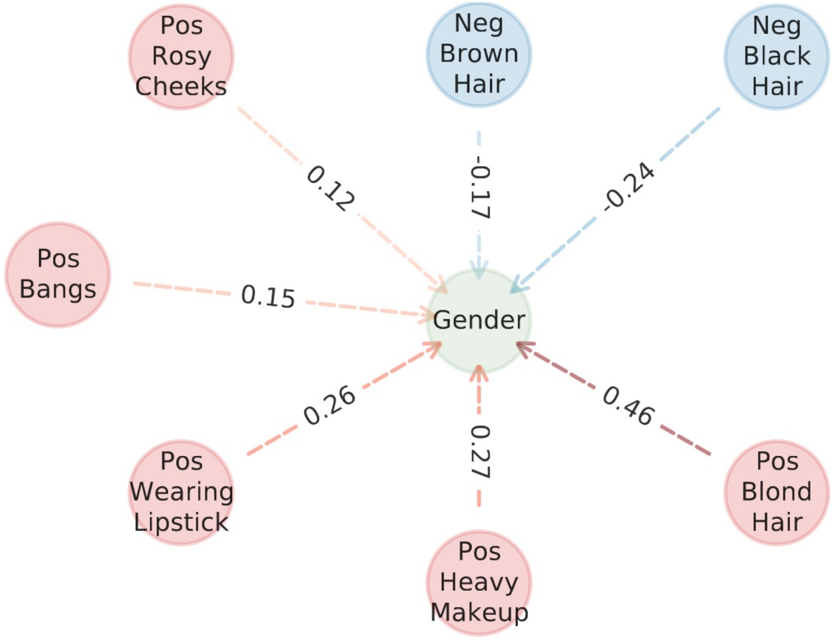
Liu, Ziwei, et al. “Deep Learning Face Attributes in the Wild.” ICCV 2015.

Discover Biases in Facial Recognition Models

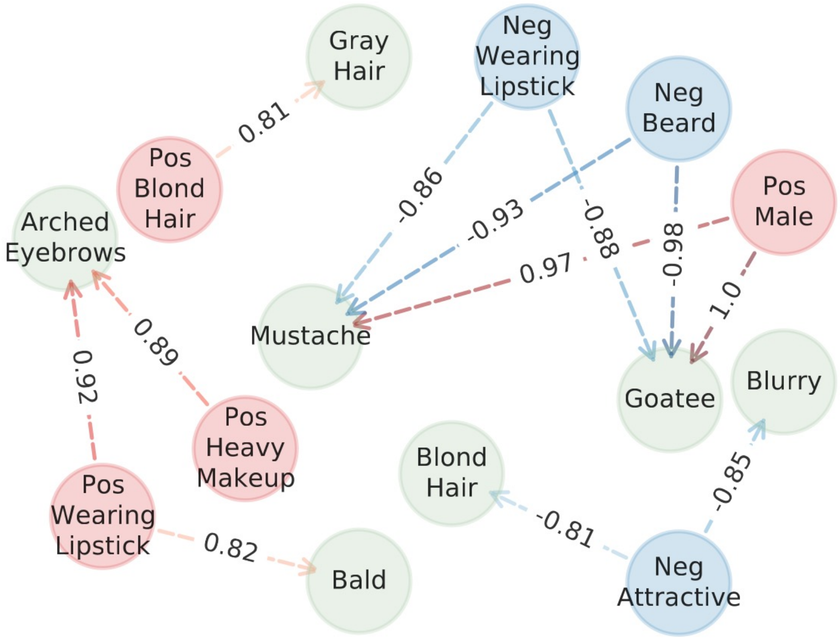


Masks distilled for predicting each attribute

Discover Biases in Facial Recognition Models

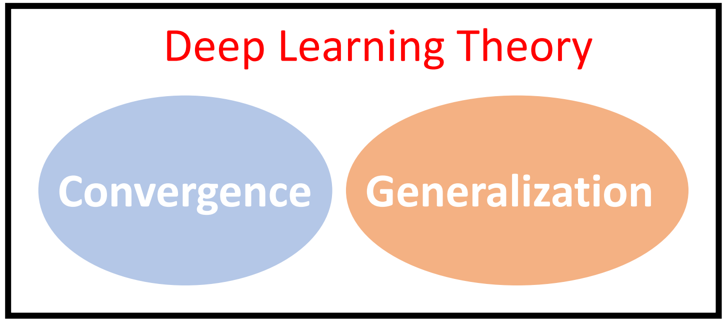


(a) Predictive attributes of *gender* attribute

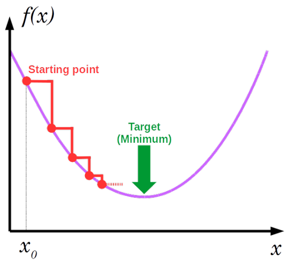


(b) All highly correlated attributes

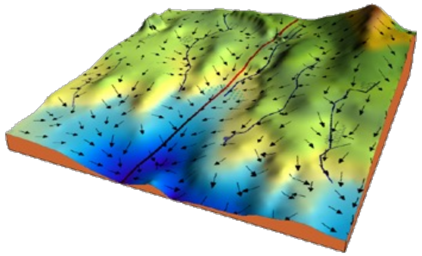
Generalization Mechanism



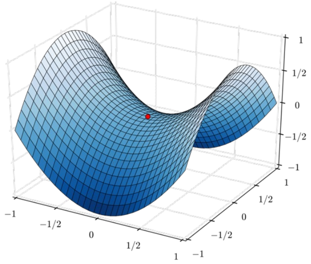
◻ Convergence



Convex (Linear model)



Nonconvex (DNN)



Saddle point

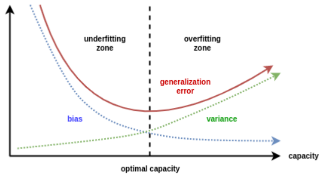
◻ Generalization

'Cat'

Training time

'Cat'?

Test time



Traditional theory: simpler model is better, more data is better

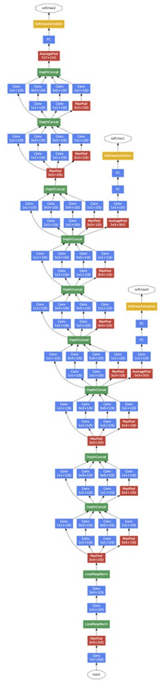
Generalization Theory

☐ Components of Generalization Error Bounds

$$\underbrace{\mathrm{err}_D(h)}_{\substack{\text{generalization}\\ \text{error}}} \;\le\; \underbrace{\widehat{\mathrm{err}}_S(h)}_{\substack{\text{empirical}\\ \text{error}}} + \underbrace{R_m(\mathcal{H})}_{\substack{\text{hypothesis}\\ \text{class}\\ \text{complexity}}} + \underbrace{\sqrt{\frac{\ln(1/\delta)}{m}}}_{\substack{\text{confidence}\\ \text{sample size}}}$$

RHS: for all terms, the lower the better:

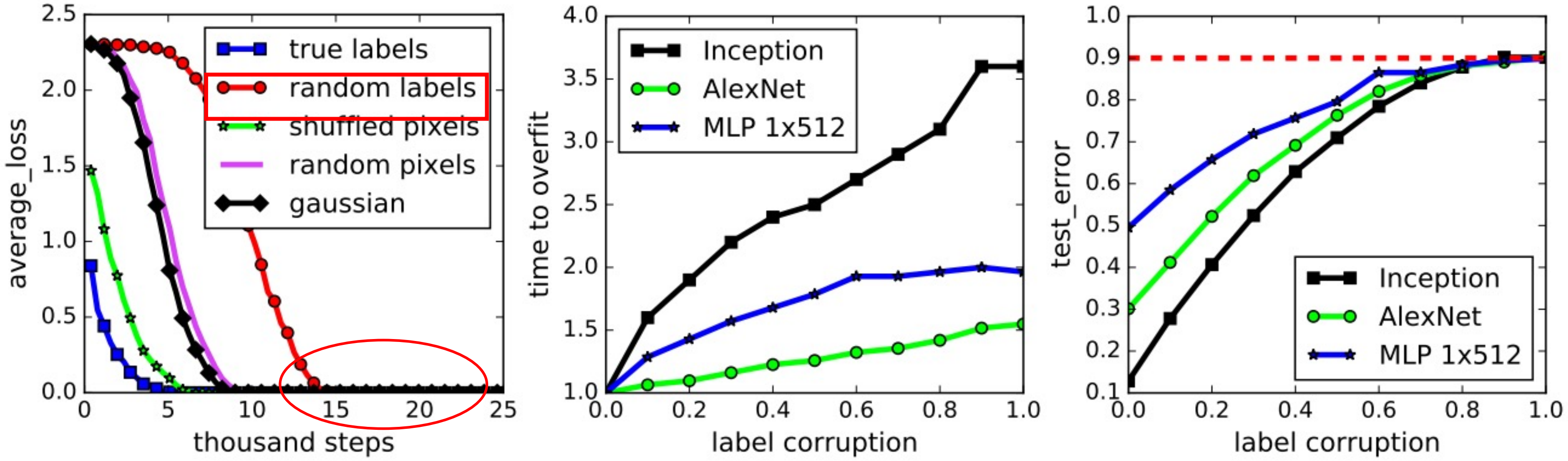
- small training error
- simpler model class
- more samples
- less confidence



https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf; https://www.youtube.com/watch?v=zlqQ7VRba2Y

Generalization Theory

☐ Small training error ≠ low generalization error



(a) learning curves (b) convergence slowdown (c) generalization error growth

Zero training error was achieved on **purely random labels** (meaningless learning)

- 0 training error vs. 0.9 test error

Zhang et al. Understanding deep learning requires rethinking generalization. ICLR 2017.

List of Existing Theories

- Rademacher Complexity bounds (Bartlett et al. 2017)
- PAC-Bayes bounds (Dziugaite and Roy 2017)
- Information bottleneck (Tishby and Zaslavsky 2015)
- Neural tangent kernel/Lazy training (Jacot et al. 2018)
- Mean-field analysis (Chizat and Bach 2018)
- Doule Descent (Belkin et al. 2019)
- Entropy SGD (Chaudhari et al. 2019)

A few interesting questions:

- ➢ Should we consider the role of data in generalization analysis?
- ➢ Should representation quality appear in the generalization bound?
- ➢ Generalization is about math (the function of the model) or knowledge?

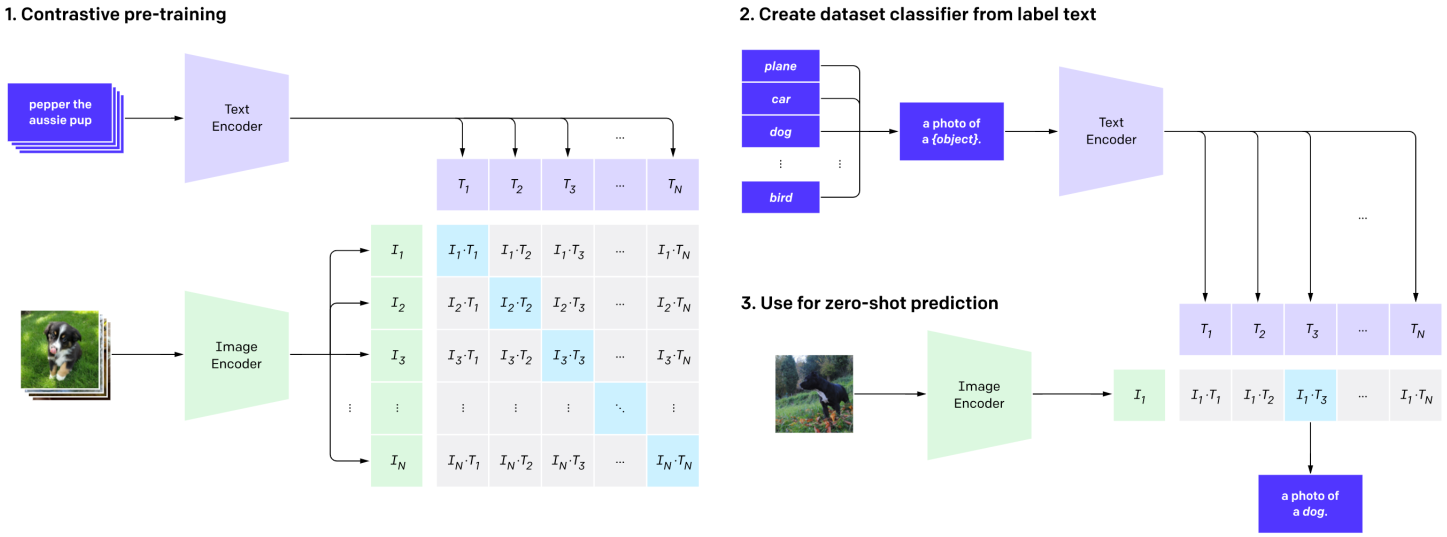
https://www.youtube.com/watch?v=zlqQ7VRba2Y

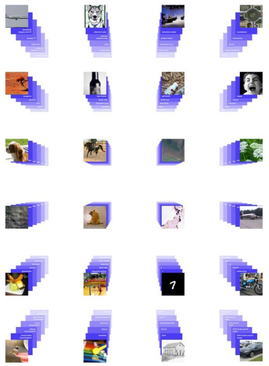
How to visualize generalization?

- Existing approaches
  - test error
  - Visualization: loss landscape, prediction attribution, etc.
  - Training -> test: distribution shift, out-of-distribution analysis
  - Noisy labels in test data – questioning data quality and reliable evaluation

- **The remaining questions:**
  - **how generalization happens?**
  - **Math ≠ Knowledge**
  - **Computation = finding patterns or understanding the underlying knowledge**
  - **What is the relation of computational generalization to human behavior?**

Cognitive Mechanism

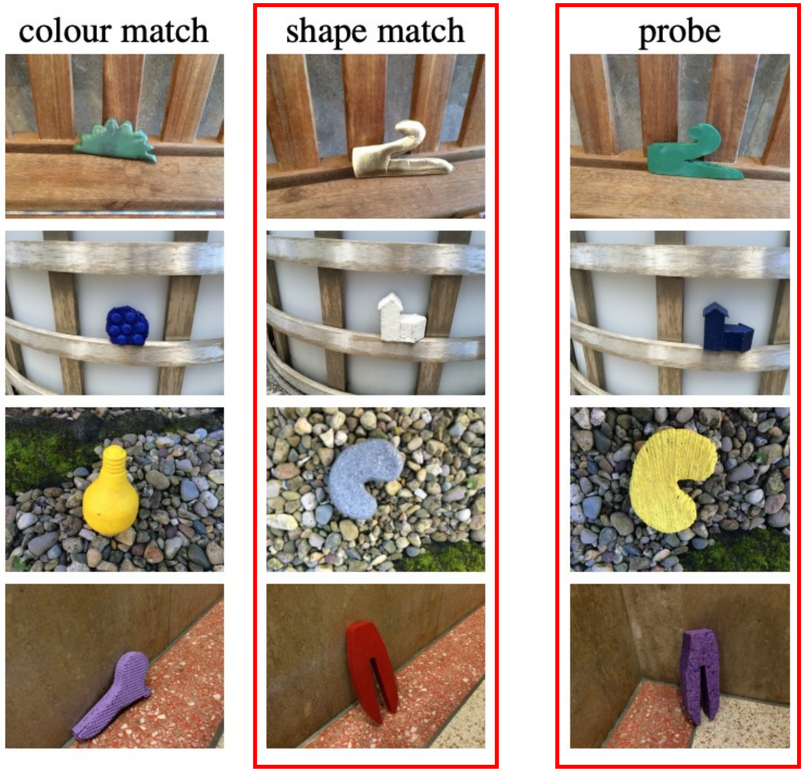




OpenAI reveals the multimodal neurons in CLIP

https://openai.com/blog/multimodal-neurons/; https://openai.com/blog/clip/

Cognitive Mechanism

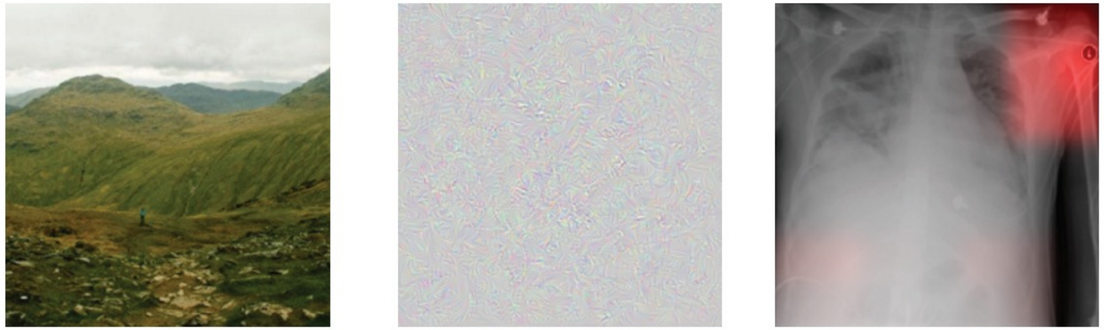


shape match = prob means **shape bias**

cognitive psychology inspired evaluation of DNNs

Ritter et al. Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study, ICML, 2017

Cognitive Mechanism



Article: Super Bowl 50

Paragraph: "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*"

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

Original Prediction: John Elway

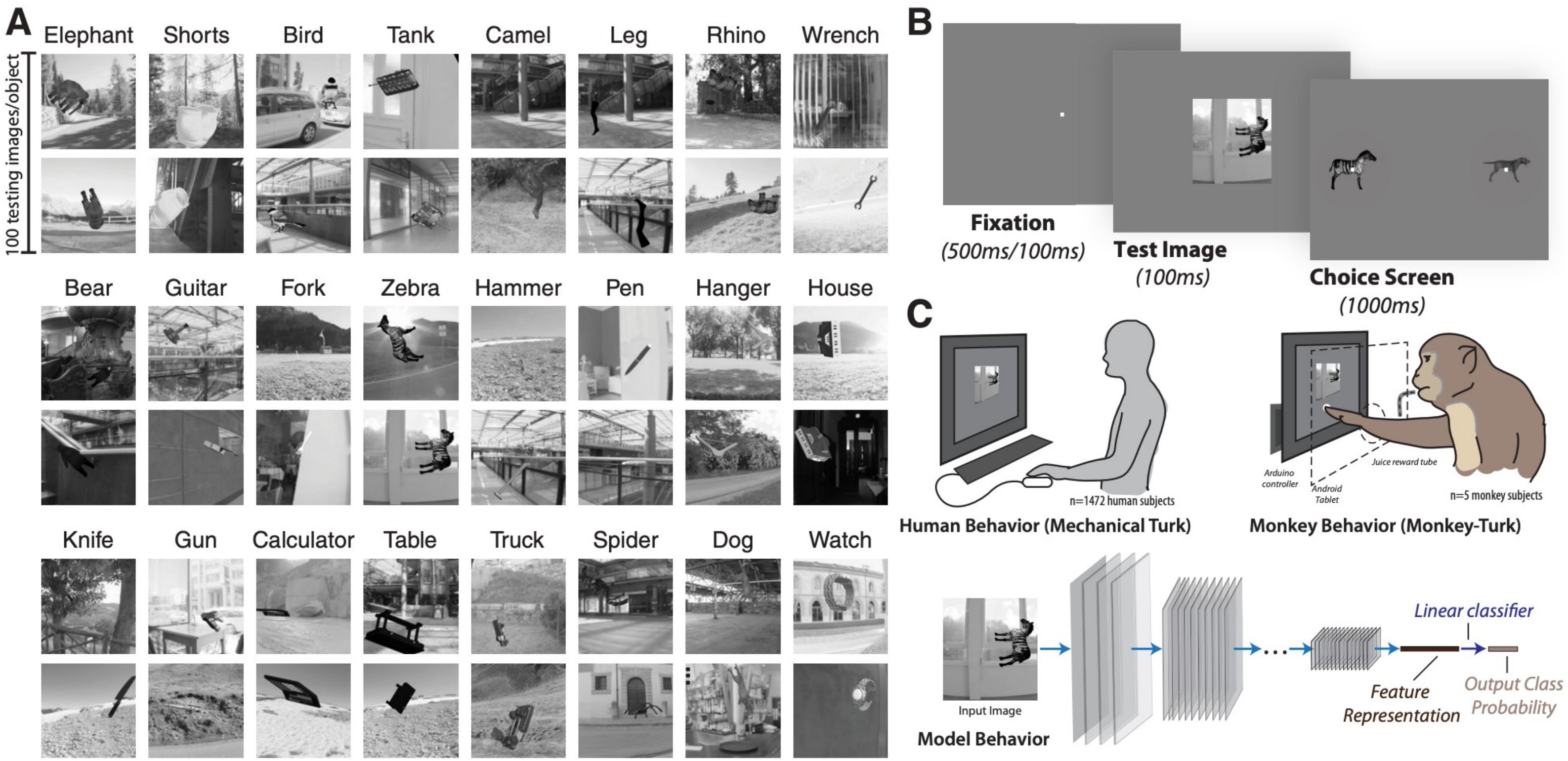
Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Deep neural networks solve problems by taking shortcuts

Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673.

Cognitive Mechanism



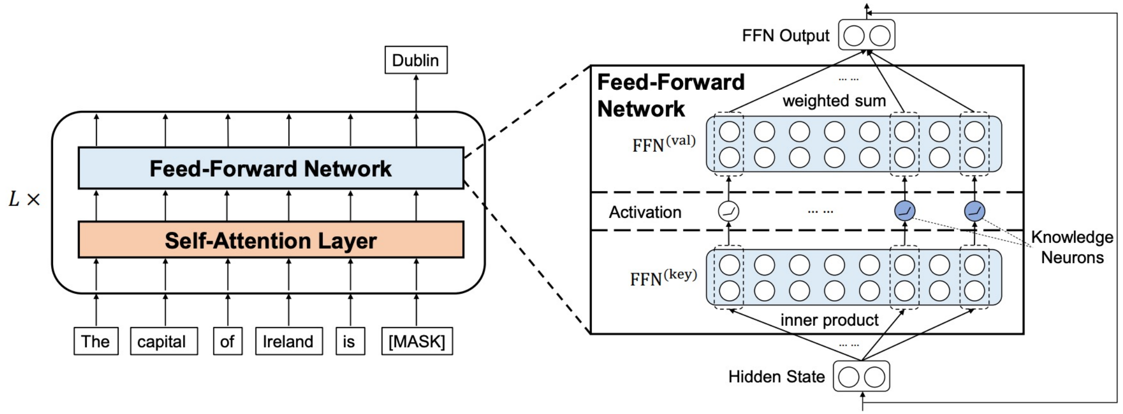
Behavioral Prediction Task: Human vs. Monkey vs. Deep Nets

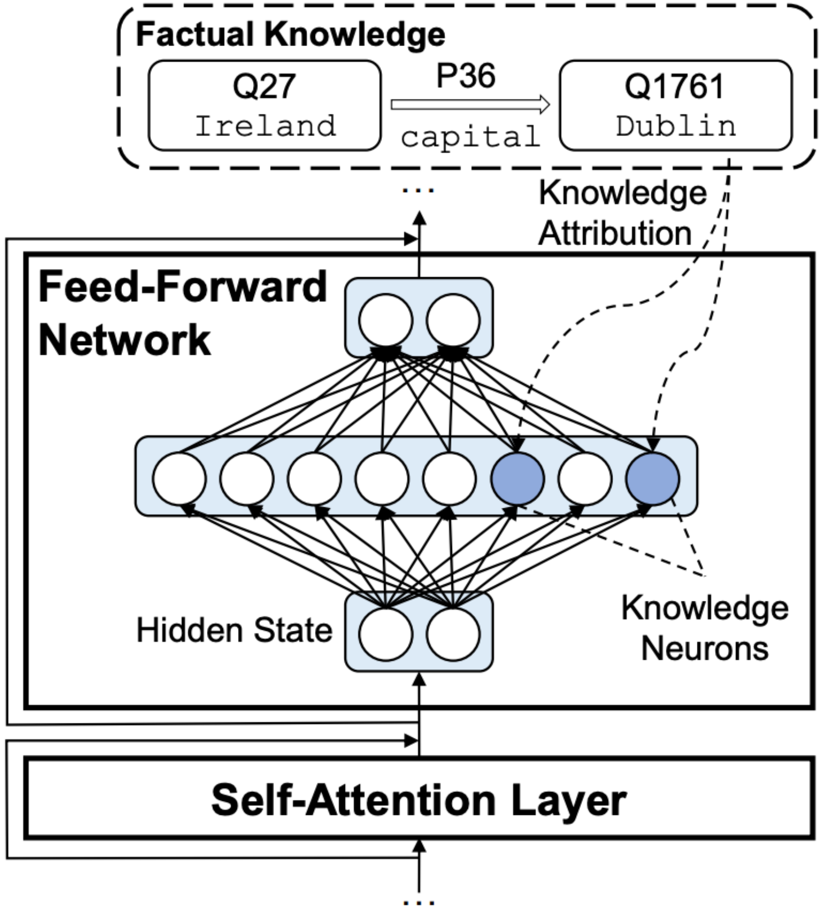
Rajalingham, Rishi, et al. “Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.” *Journal of Neuroscience* 38.33 (2018): 7255-7269. Rajalingham, Rishi, Kailyn Schmidt, and James J. DiCarlo. "Comparison of object recognition behavior in human and monkey." Journal of Neuroscience 35.35 (2015): 12127-12136.

NLP Knowledge Neurons

- Knowledge extraction/distillation
- Knowledge understanding
- Knowledge update
- Knowledge erasing

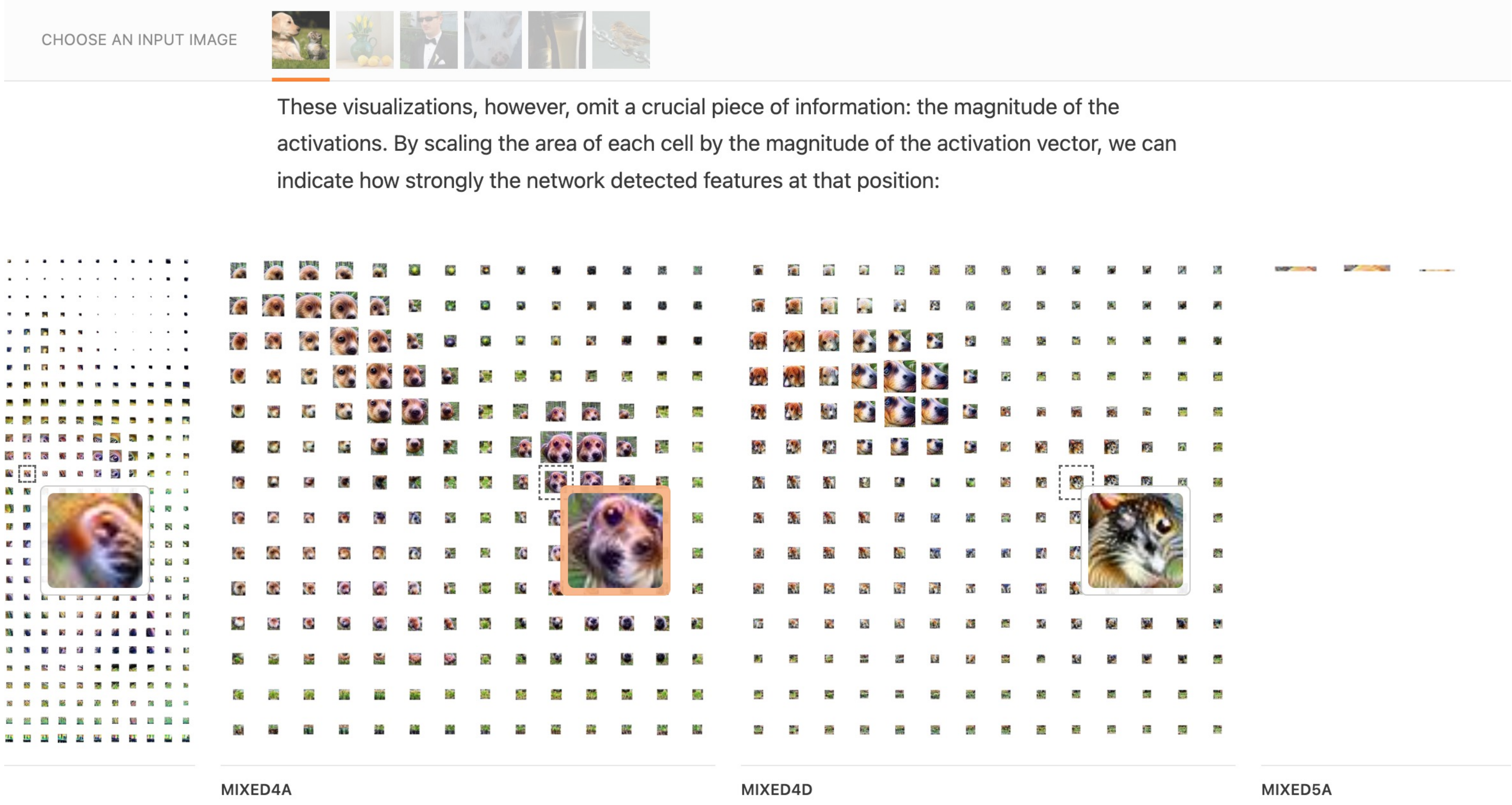
Common belief:
The FFN of a Transformer stores knowledge





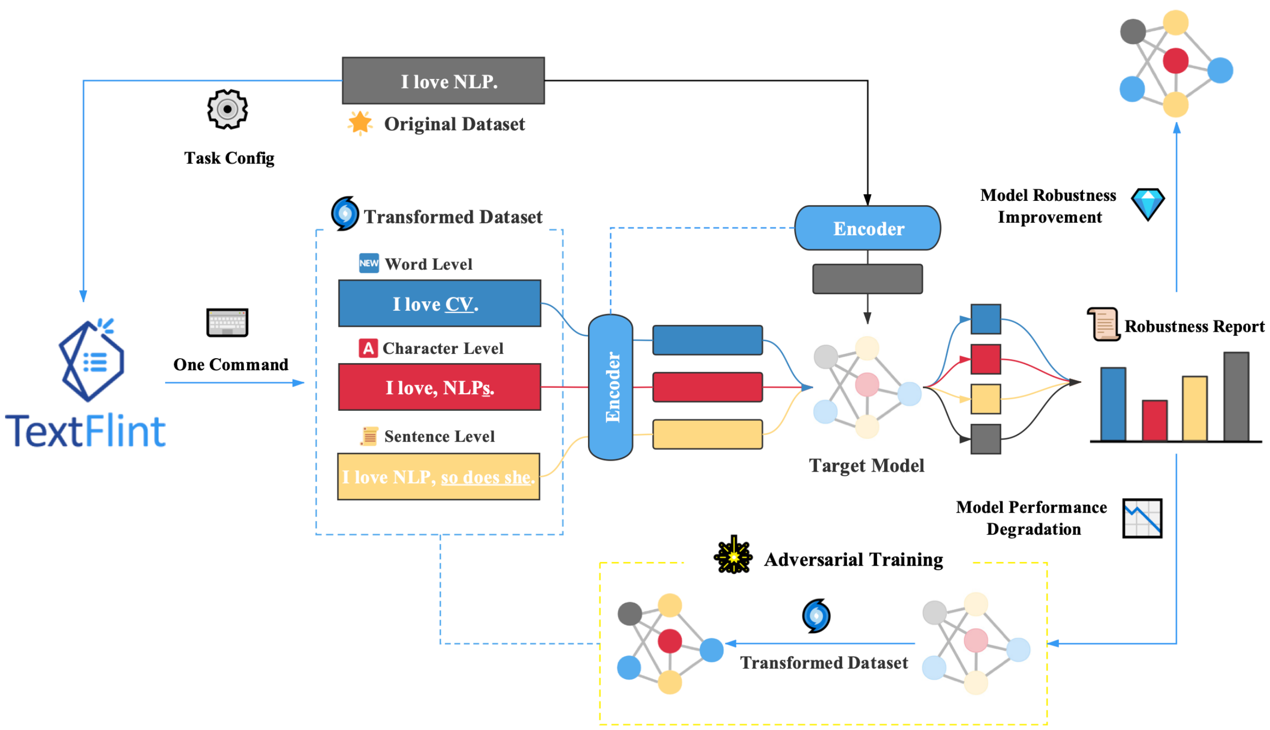
Dai, Damai, et al. "Knowledge neurons in pretrained transformers." *arXiv:2104.08696* (2021).

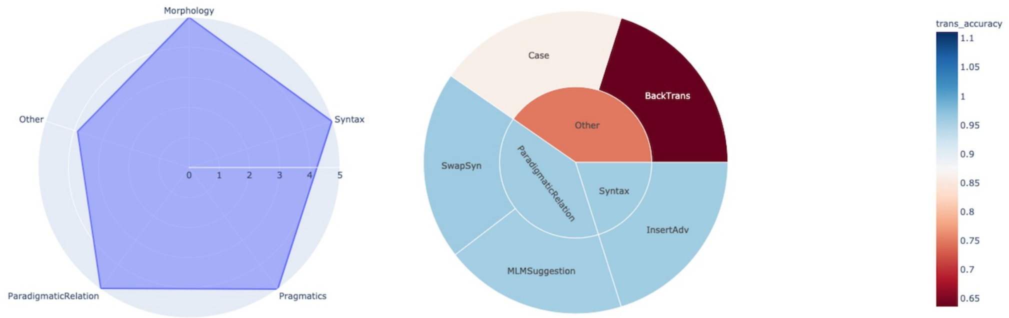
FudanNLP TextFlint

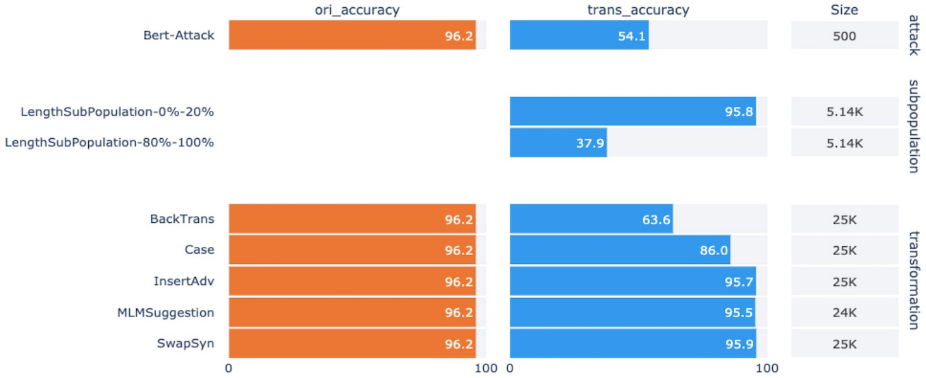


https://distill.pub/2018/building-blocks/

FudanNLP TextFlint





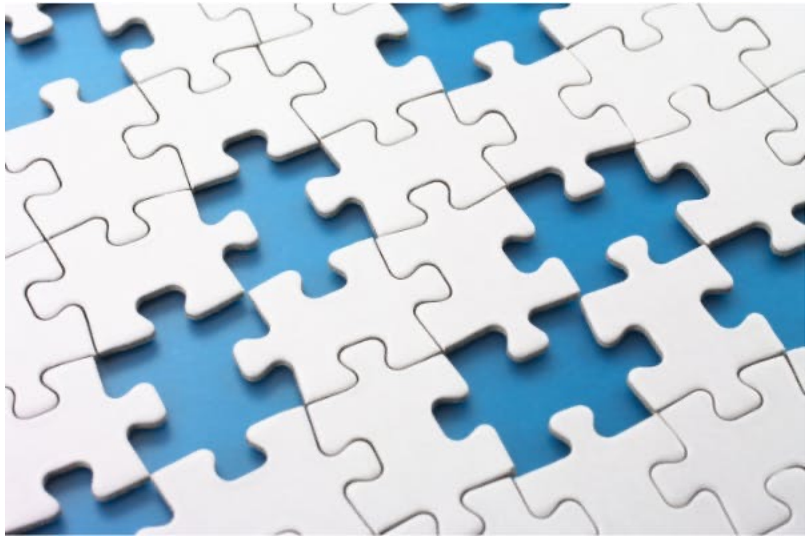


https://textflint.github.io/; https://github.com/textflint/textflint

What is Missing

Many theoretical work or interpretation tools have been proposed

Yet, we don’t have an all-in-one system to explain everything.

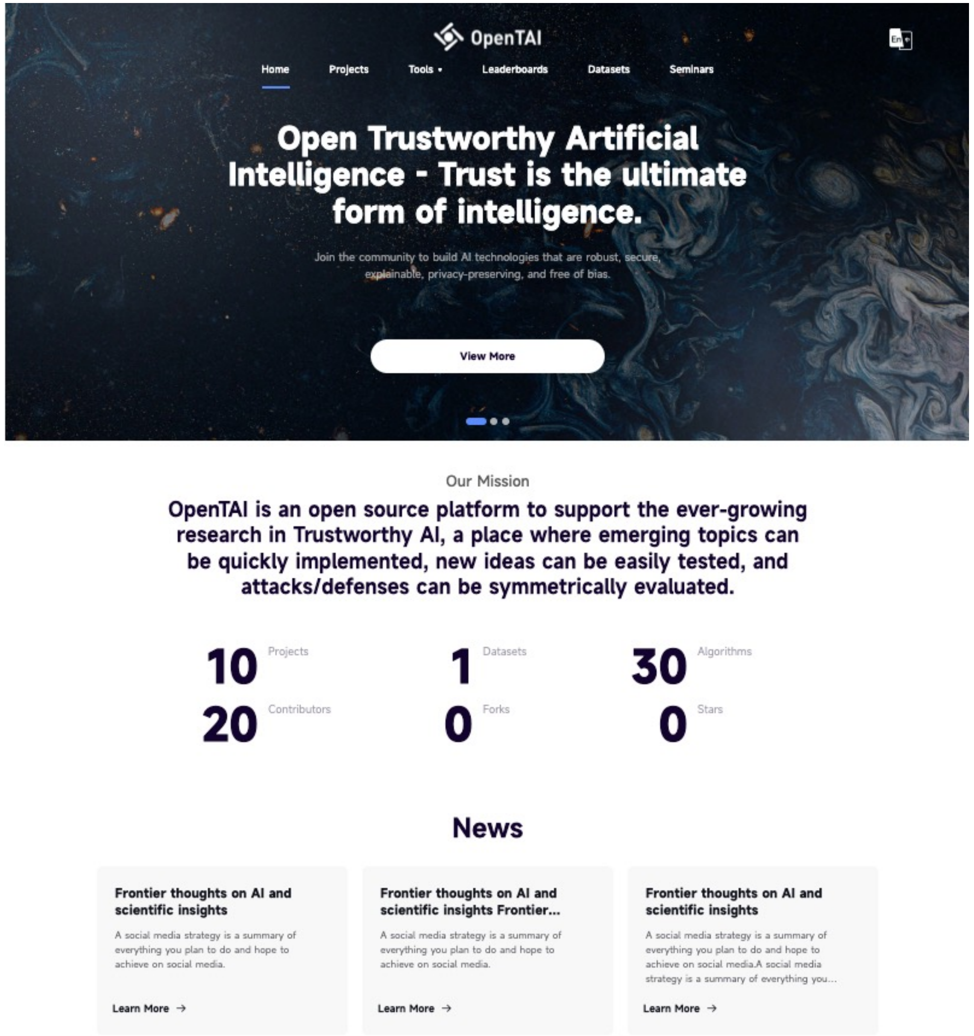


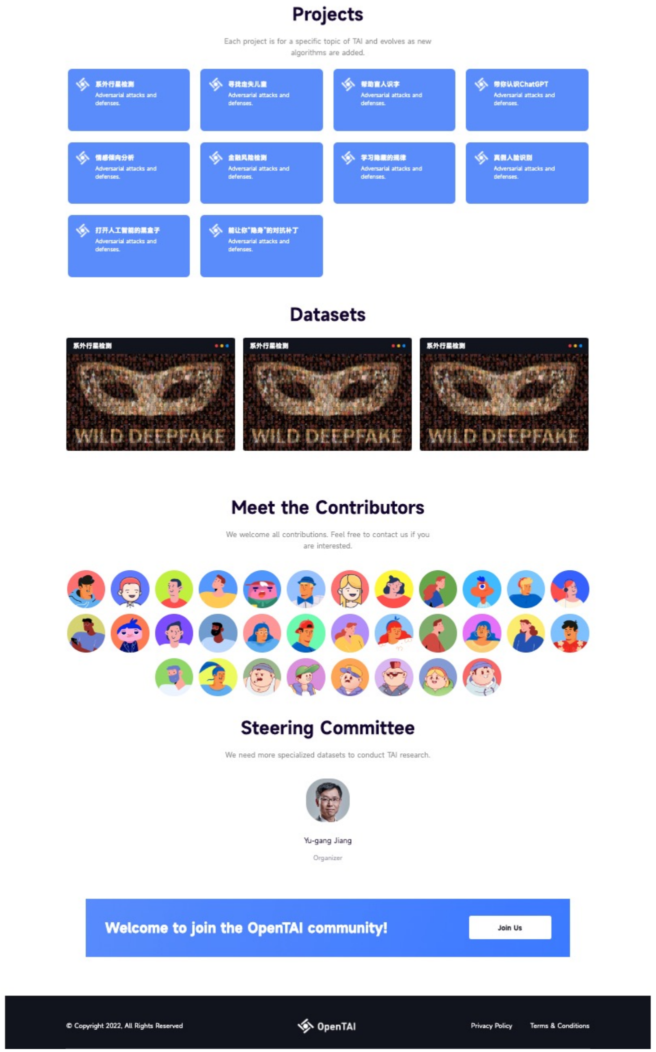
AI治理开放平台+攻击检测工具集

- 与浦江实验室和清华大学共同发布 **“蒲公英”人工智能治理开放平台**，积极应对AI鲁棒性问题和全球治理挑战

2022年世界人工智能大会
科学前沿全体会议公开发布

开放可信AI社区 (OpenTAI)





https://opentai.org/

攻击展示

系统分析展示AI可信与安全性问题：

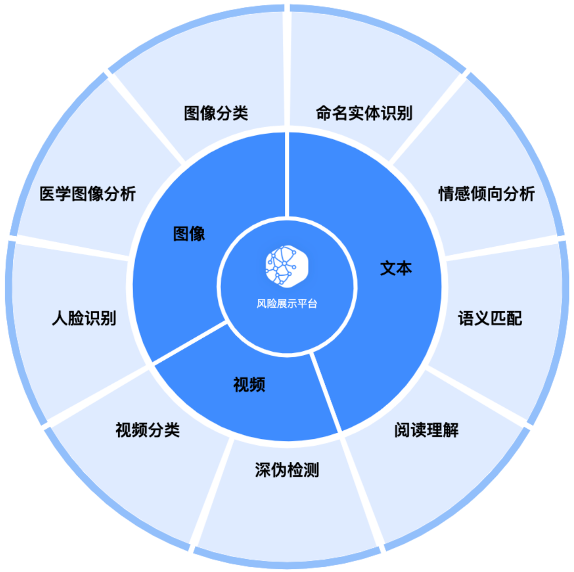
- **3种媒体**：图像、视频、文本
- **9大任务**：图像分类、医学图像分析、人脸识别、视频分类、深伪检测、命名实体识别、情感倾向分析、语义匹配、阅读理解
- **36个模型**：ResNet、Transformer等
- **6大维度**：性能、安全性、鲁棒性、可解释性、隐私性、公平性

风险展示平台由一套交互式界面和风险分析工具组成。旨在帮助大众和决策制定者理解当前人工智能模型所存在的风险，以及风险所触发的伦理治理规则，发展规则和技术互促的人工智能治理研究。目前，平台支持36种常见视觉和语言模型的交互分析，其中包括图像分类、医学图像分析、情感倾向分析、命名实体识别、语义匹配、阅读理解、视频分类、深伪检测等模型。

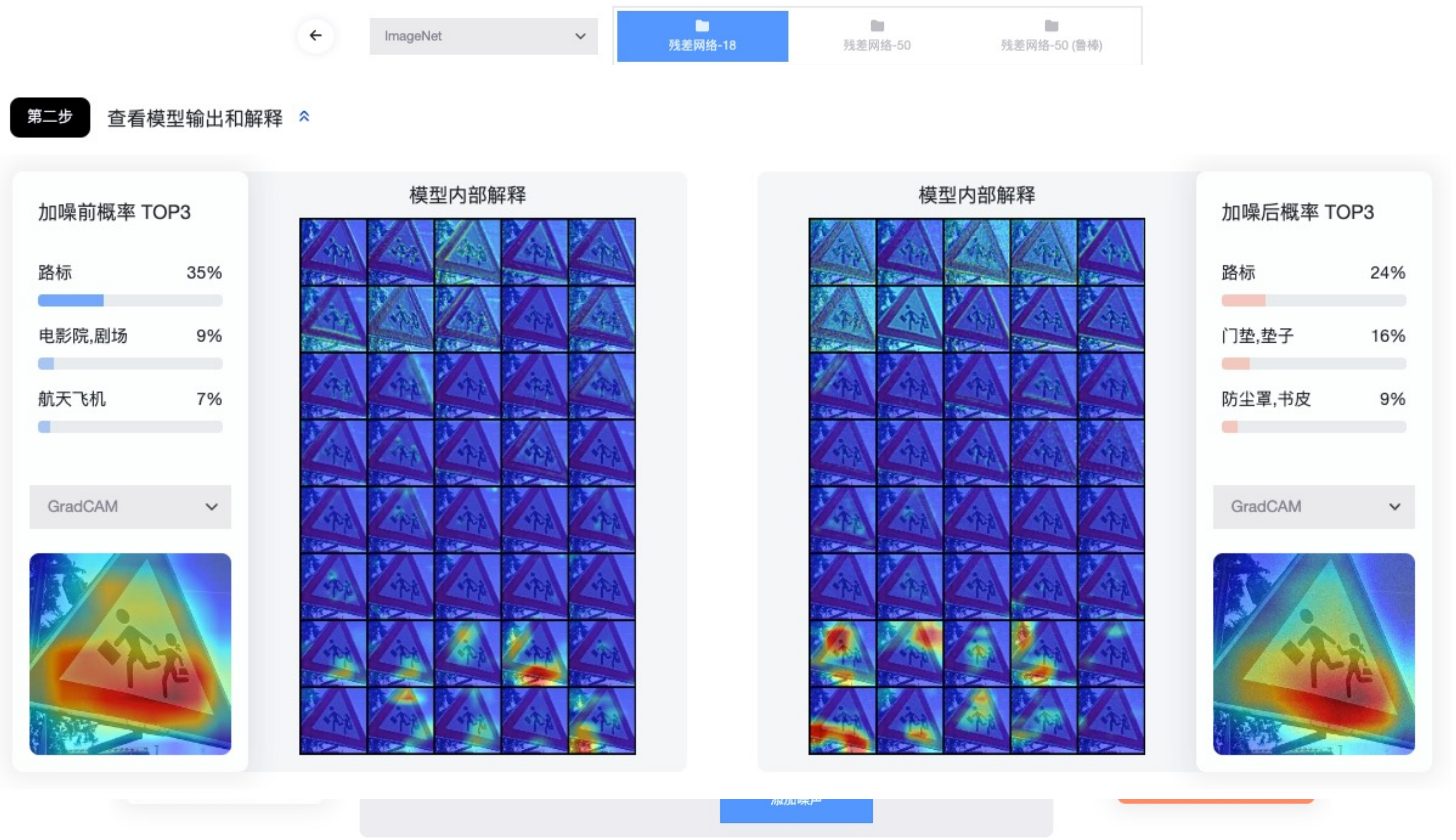
支持的交互分析包括基于普通噪声的鲁棒性分析、基于对抗攻击技术的安全性分析以及基于数据窃取技术的隐私性分析。我们将持续建设此平台，逐步增加更多的应用场景、模型和分析工具。

第一步 选择样本 → 第二步 添加噪声 → 第三步 执行模型 → 第四步 查看结果

请选择要查看的任务类别

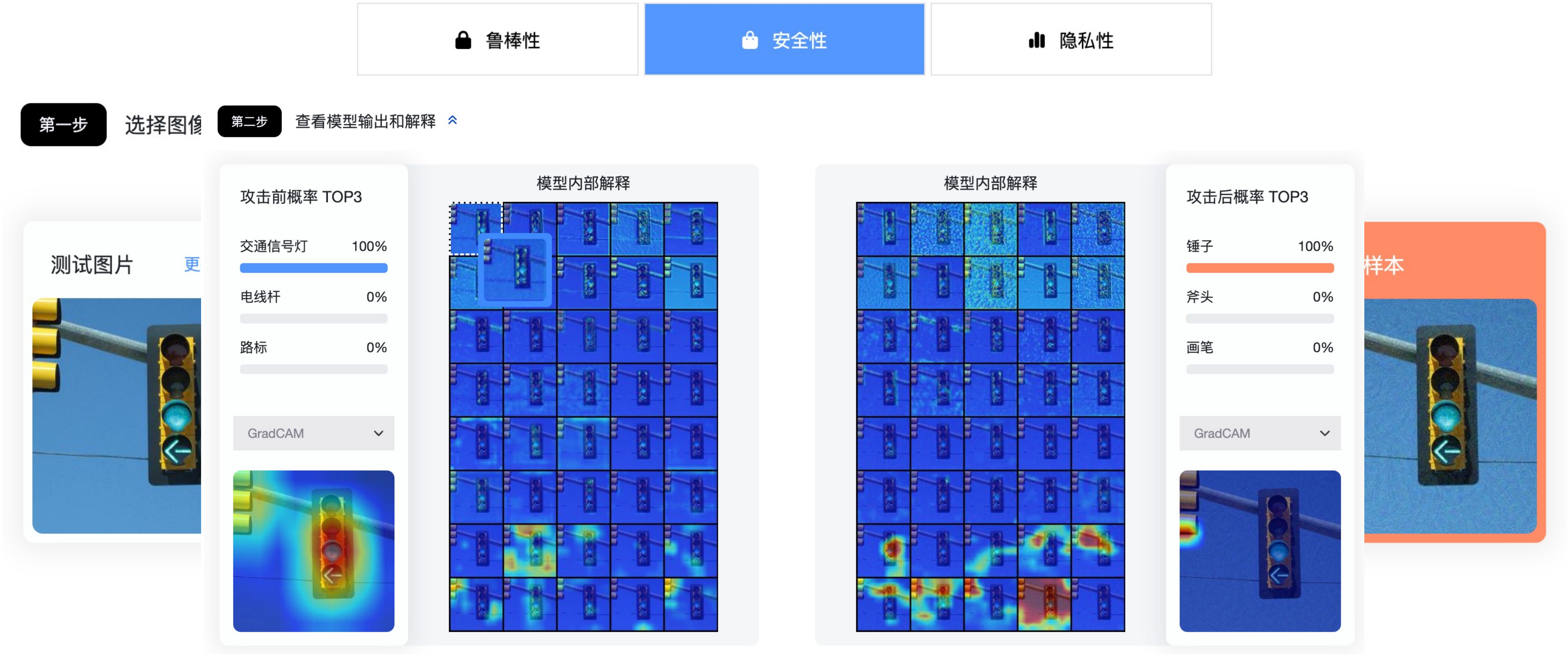


举例 – 图像分类



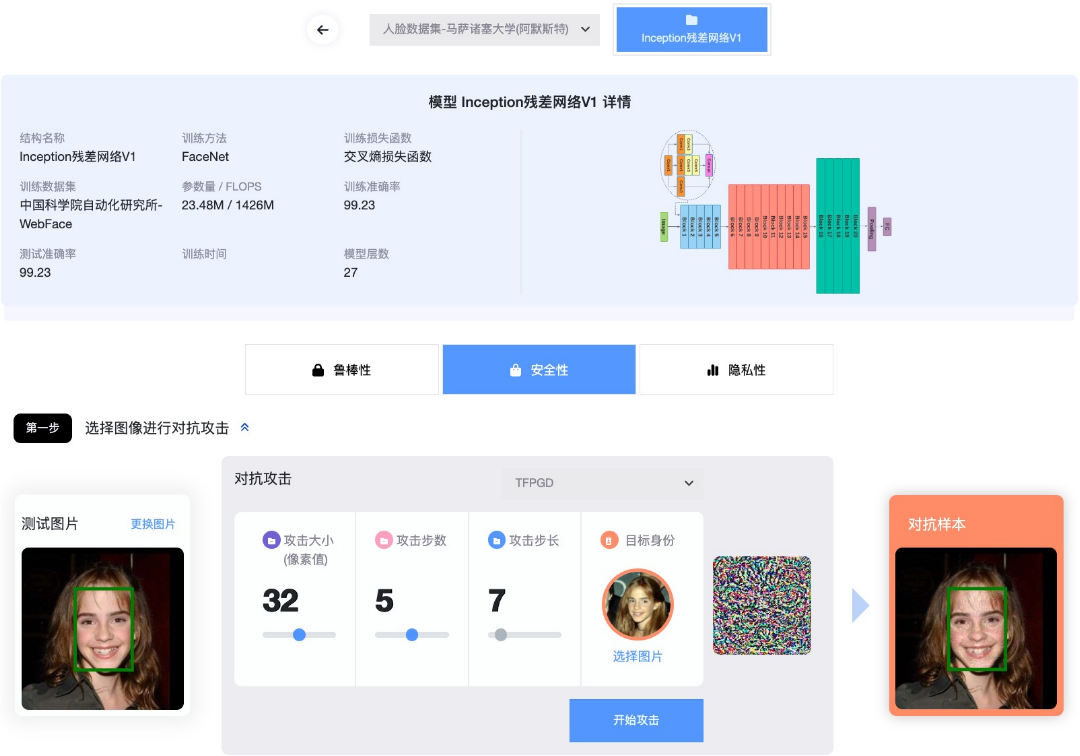
ImageNet
残差网络-18
残差网络-50
残差网络-50 (鲁棒)
第二步
查看模型输出和解释
加噪前概率 TOP3
路标 35%
电影院,剧场 9%
航天飞机 7%
GradCAM
模型内部解释
模型内部解释
加噪后概率 TOP3
路标 24%
门垫,垫子 16%
防尘罩,书皮 9%
GradCAM

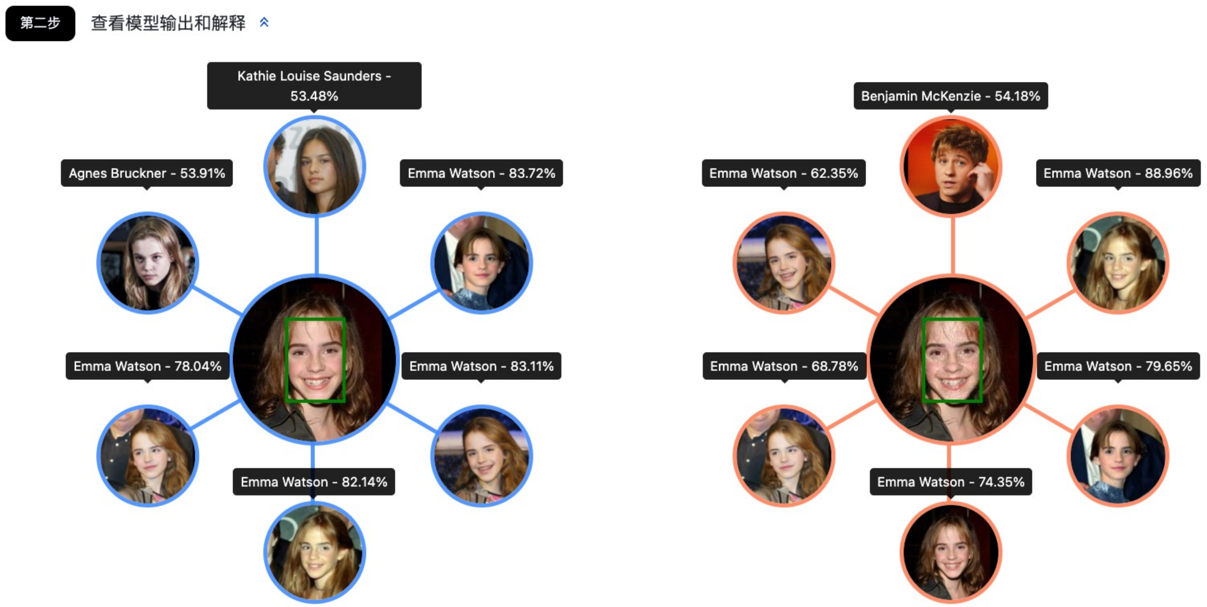
举例 – 图像分类



https://opentai.org/

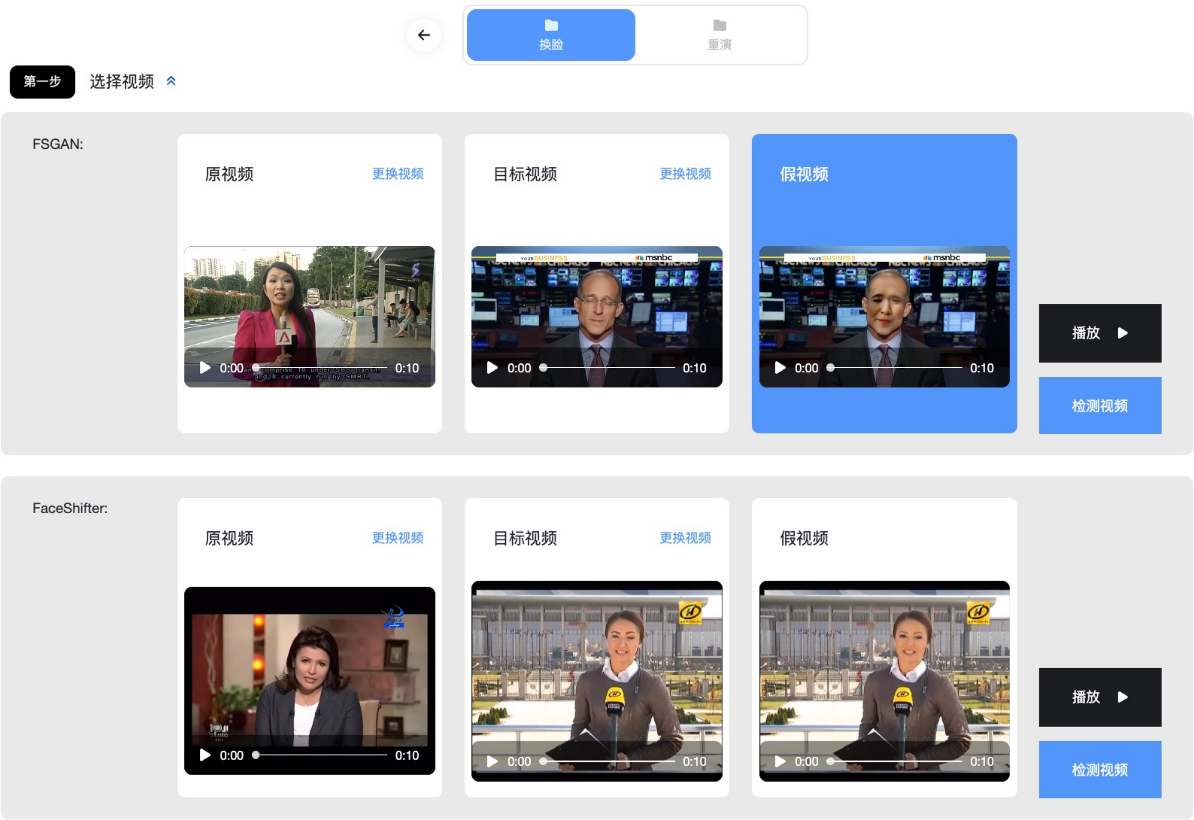
举例 – 人脸识别



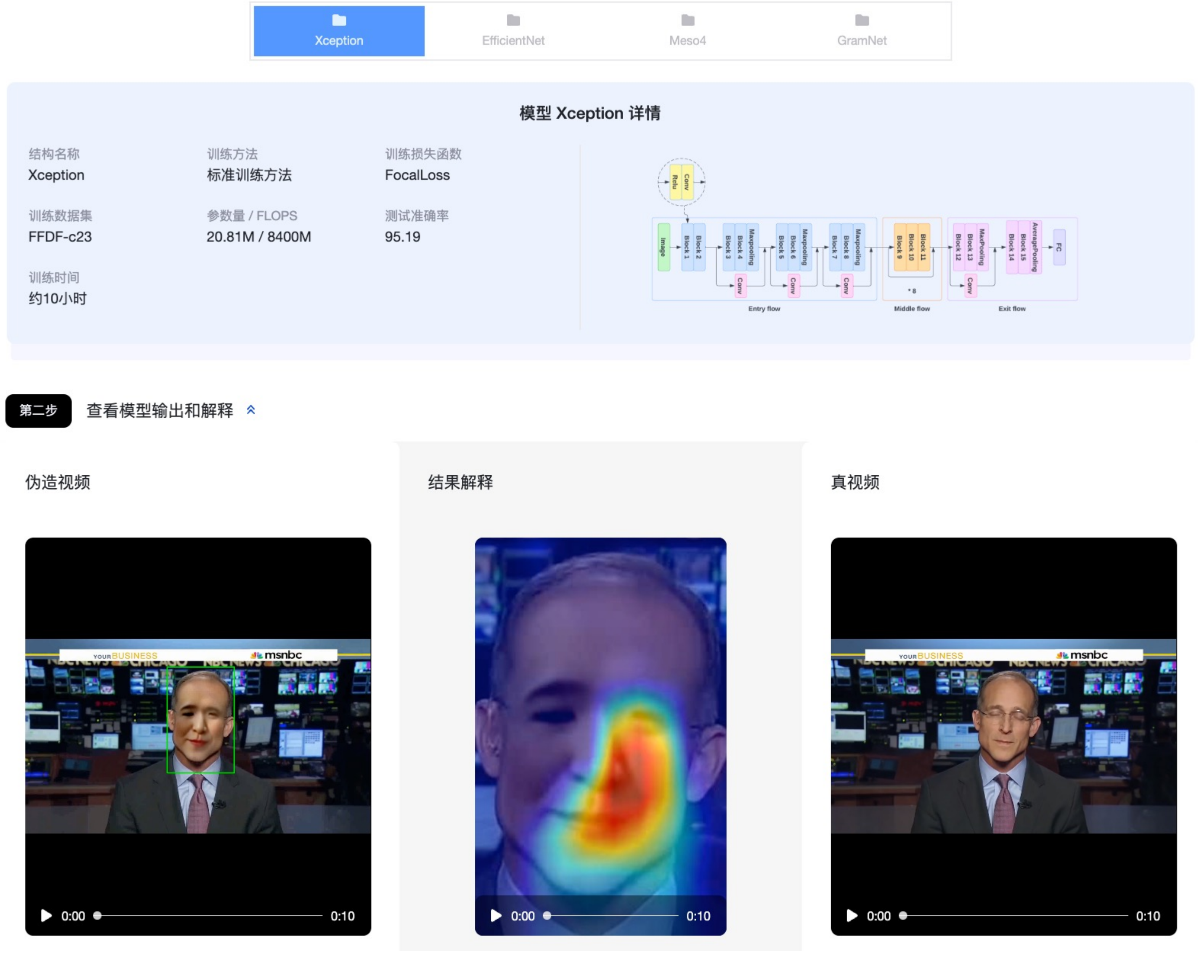


https://opentai.org/

举例 – 深度伪造检测



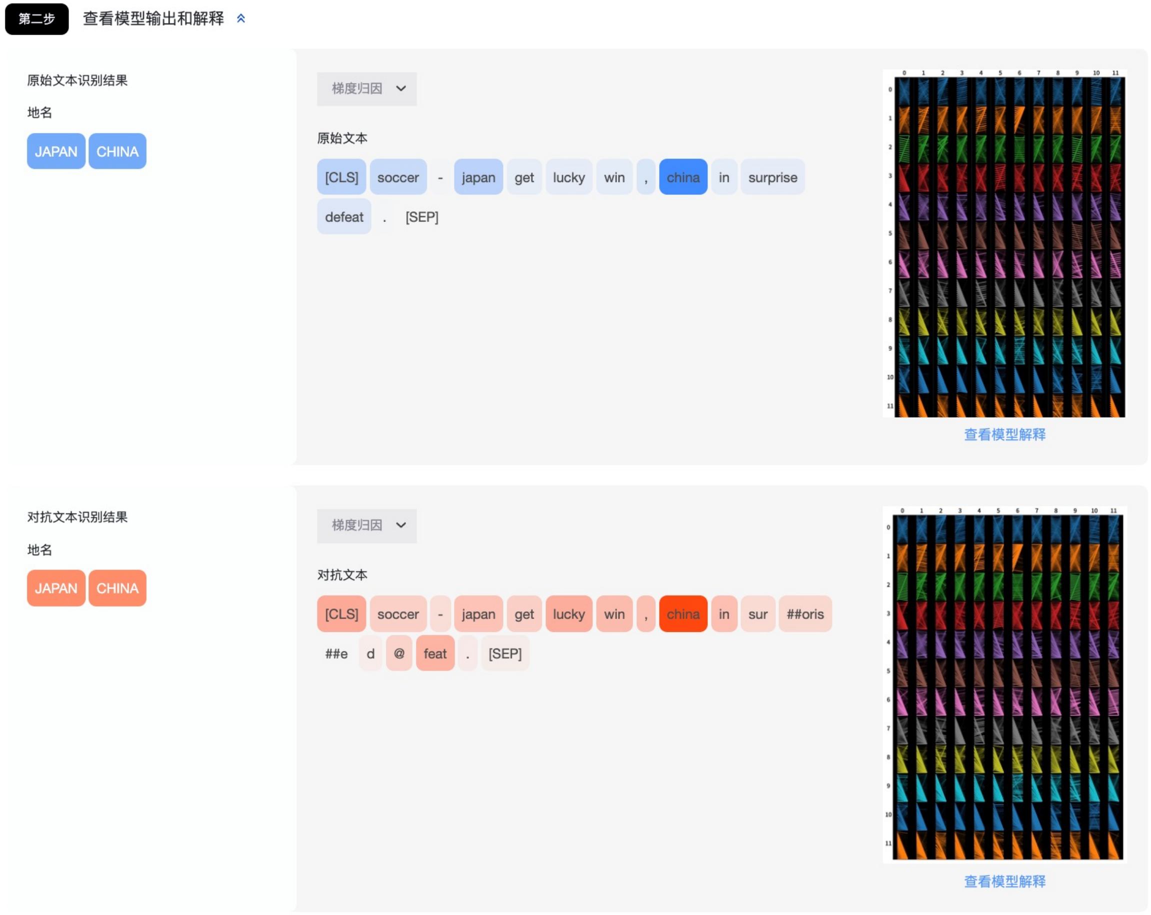
换脸
重演
第一步
选择视频
FSGAN:
原视频
更换视频
目标视频
更换视频
假视频
播放
检测视频
FaceShifter:
原视频
更换视频
目标视频
更换视频
假视频
播放
检测视频



Xception
EfficientNet
Meso4
GramNet
模型 Xception 详情
结构名称
Xception
训练方法
标准训练方法
训练损失函数
FocalLoss
训练数据集
FFDF-c23
参数量 / FLOPS
20.81M / 8400M
测试准确率
95.19
训练时间
约10小时
第二步
查看模型输出和解释
伪造视频
结果解释
真视频

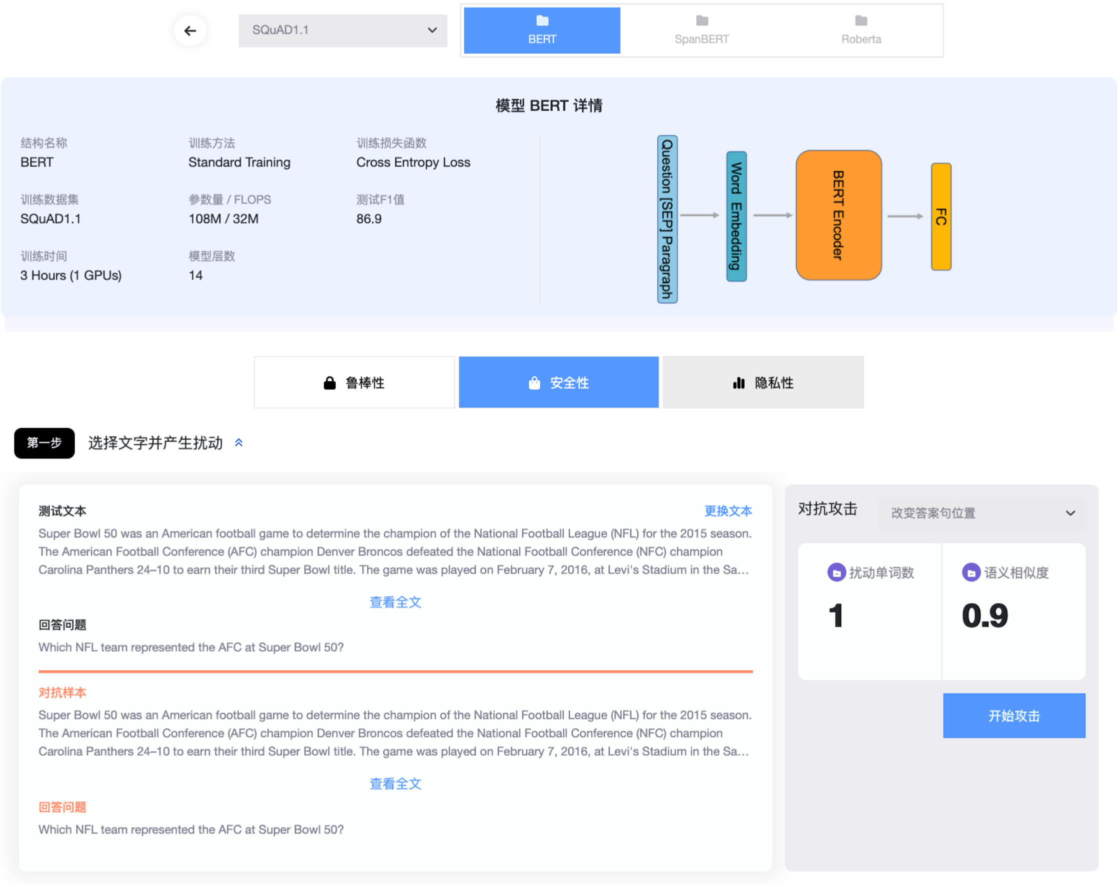
举例 – 命名实体识别

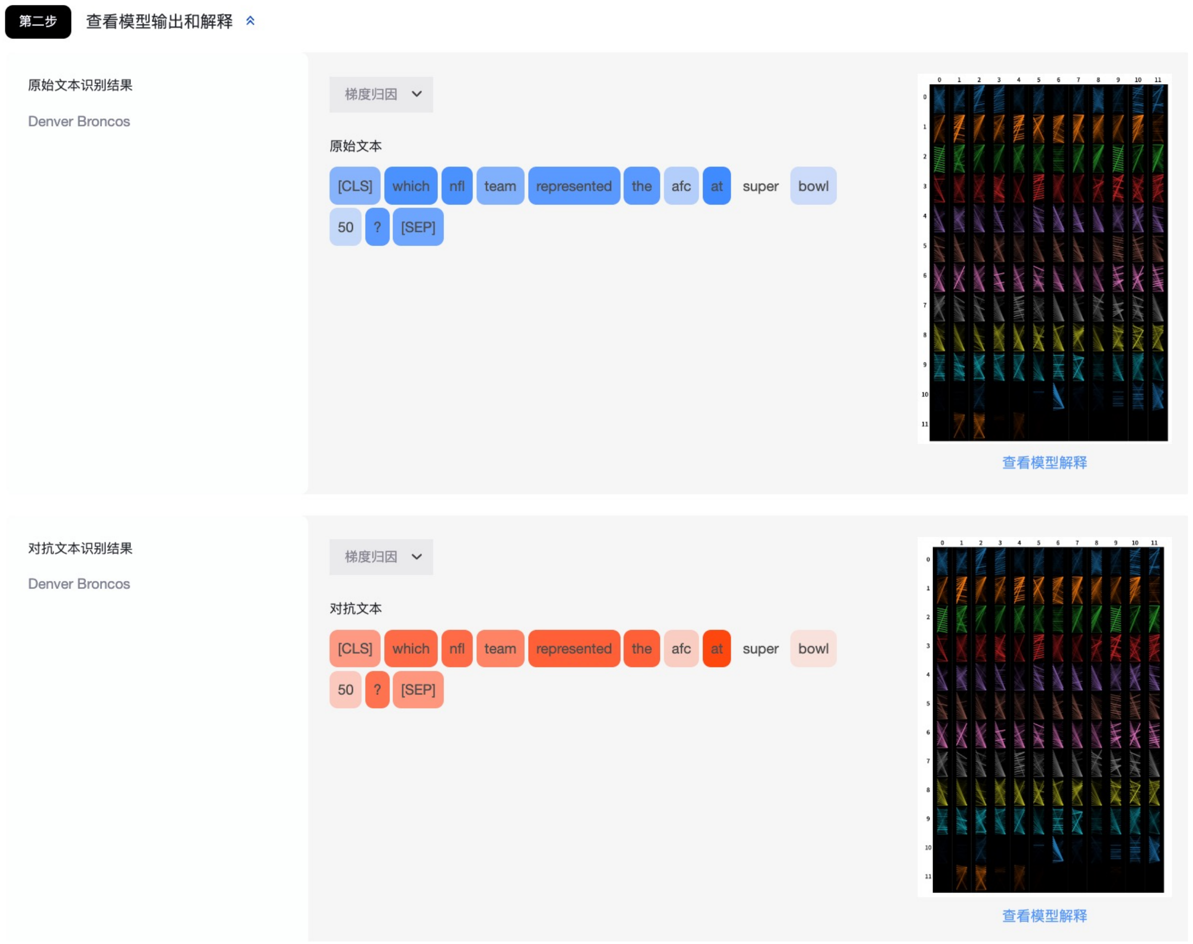
CoNLL2003
Bert拼接条件随机场
基于transformer的实体识别网络
基于信息论的命名实体识别
模型 Bert拼接条件随机场 详情
结构名称
Bert拼接条件随机场
训练方法
标准训练方法
训练损失函数
交叉熵损失函数
训练数据集
CoNLL2003
参数量 / FLOPS
108.32M / 32M
训练F1值
99.8
测试F1值
90.41
训练时间
单GPU 2 小时
模型层数
14
Text
Embedding
Transformer 1
Transformer 2
Transformer 3
Transformer 4
Transformer 5
Transformer 6
FC
CRF
鲁棒性
安全性
隐私性
第一步
选择文字并产生扰动
测试文本
更换文本
SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
噪声样本
SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURORISE D@FEAT .
普通噪声
字符扰动噪声
扰动单词数
1
语义相似度
0.9
添加噪声



第二步
查看模型输出和解释
原始文本识别结果
地名
JAPAN
CHINA
梯度归因
原始文本
[CLS] soccer - japan get lucky win , china in surprise defeat . [SEP]
查看模型解释
对抗文本识别结果
地名
JAPAN
CHINA
梯度归因
对抗文本
[CLS] soccer - japan get lucky win , china in sur ##oris ##e d @ feat . [SEP]
查看模型解释

举例 – 阅读理解





https://opentai.org/

举例 – 模型逆向/数据窃取

第二步 查看模型输出和解释

逆向图像 -> 从模型中通过逆向技术窃取出来的数据

原始图片 -> 模型训练所使用的原始数据

https://opentai.org/

评估评测

模型鲁棒性评测：

- **2种媒体**：图像、文本
- **6种任务**：图像分类、医学图像分析、命名实体识别、情感倾向分析、语义匹配、阅读理解
- **20个模型**：ResNet、Transformer等
- **6大维度**：性能、安全性、鲁棒性、可解释性、隐私性、公平性

六大维度评估体系

1. **性能**：模型在测试数据集上的性能
2. **鲁棒性**：模型对多种常见噪声的鲁棒性
3. **安全性**：模型对常见对抗攻击的安全性
4. **可解释性**：可解释部分信息占总信息的比率
5. **隐私性**：模型逆向所泄露的信息占总信息的比率
6. **公平性**：模型在两组对照测试数据上的表现差异

模型评分排行

1. **排行范围**：服务于同种任务的不同模型
2. **排行依据**：六大维度平均得分
3. **评分方式**：各模型独立评估计算得分
4. **评分标准**：独创 0 – 5 级评分标准

选择应用场景

图像	图像分类	医学图像分析		
文本	命名实体识别	情感倾向分析	语义匹配	阅读理解

举例 – 医学图像分类模型

六维评测

模型评分详情

性能 ★★★★
在1000张测试图片上达到了82.95%的top-1分类准确率。

鲁棒性 ★★★★
在少量普通噪声干扰下能够保持61.55%的top-1分类准确率。

安全性 ★★
在少量对抗噪声干扰下能够保持13.58%的top-1分类准确率。

可解释性 ★★★★
在推理阶段，遮挡可解释信息后（先验），模型的概率平均下降86.99%，与真实可解释信息吻合度高。

隐私性 ★★★
在数据窃取攻击下，模型会泄漏50.50%的信息。

公平性 ★★★★
在多组公平性对比测试图片上，模型决策一致性为81.81%。

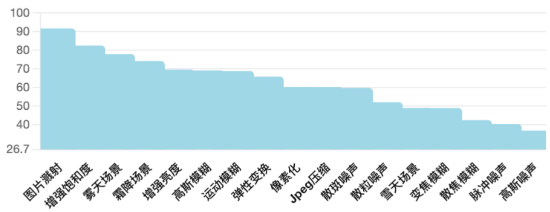
模型评分排行

1	残差网络-50 (鲁棒)	3.92 分
2	残差网络-50	3.42 分

1 性能

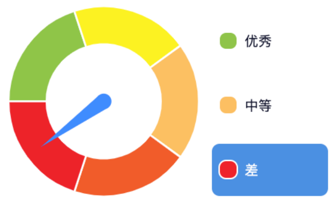
性能 解释
在1000张测试图片上达到了82.95%的top-1分类准确率，最准确的类别是92.05%，最差的类别是73.86%。

2 鲁棒性



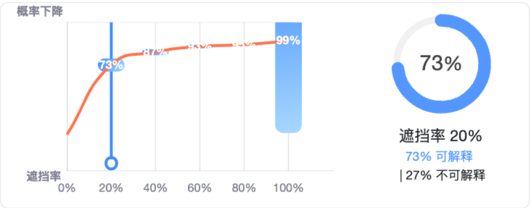
鲁棒性 解释
在1000张测试图片上进行了17种普通噪声测试，在少量噪声干扰下，准确率平均下降21.40%。

3 安全性



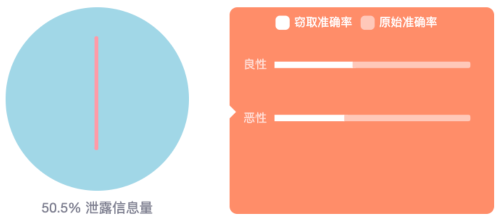
安全性 解释
在1000张测试图片上，使用对抗攻击方法PGD, 在扰动氛围1, 2, 3, 4, 6, 8像素下，5种强度的攻击，模型准确率下降'69.37%', '73.98%', '77.27%', '80.91%', '79.93%', '81.64%', 平均下降为77.18%。

4 可解释性



可解释性 解释
在1000张测试图片上，进行可解释（GradCam）信息遮挡测试，不同遮挡比例会导致不同程度的概率下降。对于ImageNet数据集，假设先验知识是"图像中20%的区域是跟分类物体相关的（20%的信息是可解释的）那么遮挡20%的关键信息后，概率下降越多可解释性越好。

5 隐私性



隐私性 解释
对模型的2个类别进行数据逆向尝试，然后再逆向后的数据上进行模型训练，得到的新模型的性能代表了信息泄漏的多少。原模型可能会泄露50.50%的信息，其中泄漏最多的是良性类别，为60.51%。

6 公平性

公平性 解释
类间不公平性：选择两组（类）数据进行对比实验，如果模型在这两组数据上预测置信度一致，则结果是公平的，否则结果差异越大，公平性越差。

A little bit more on: Common Robustness

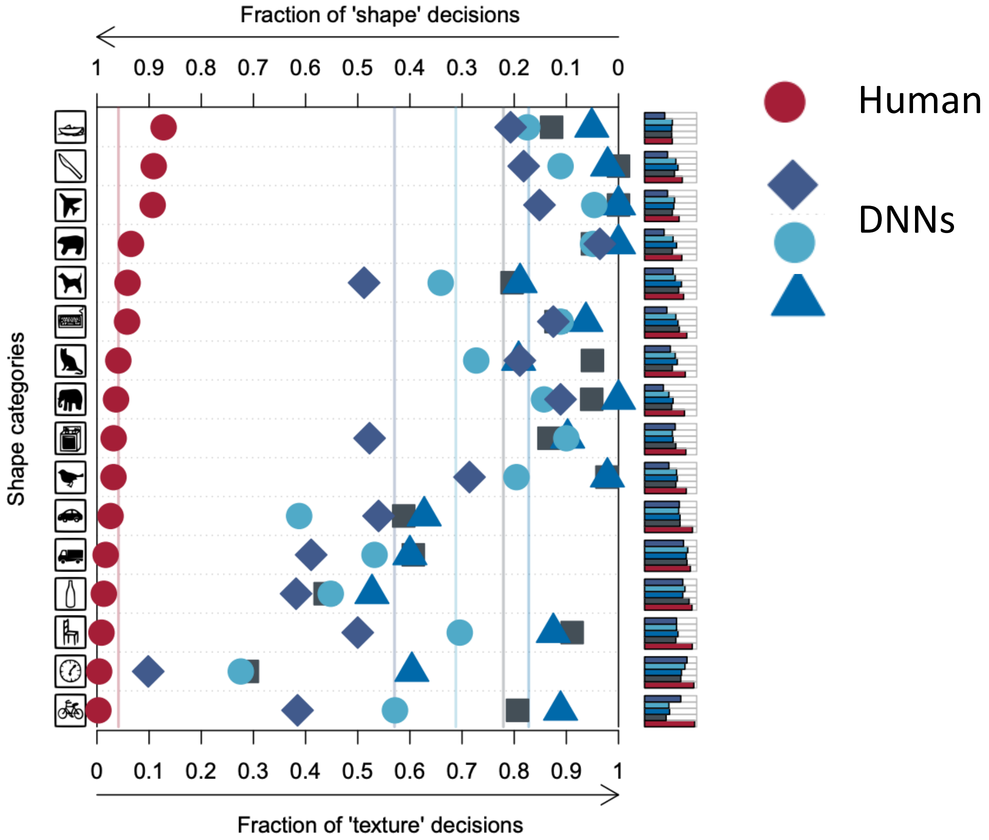
- **Texture bias**
- **Robustness to common corruptions**

Texture bias

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

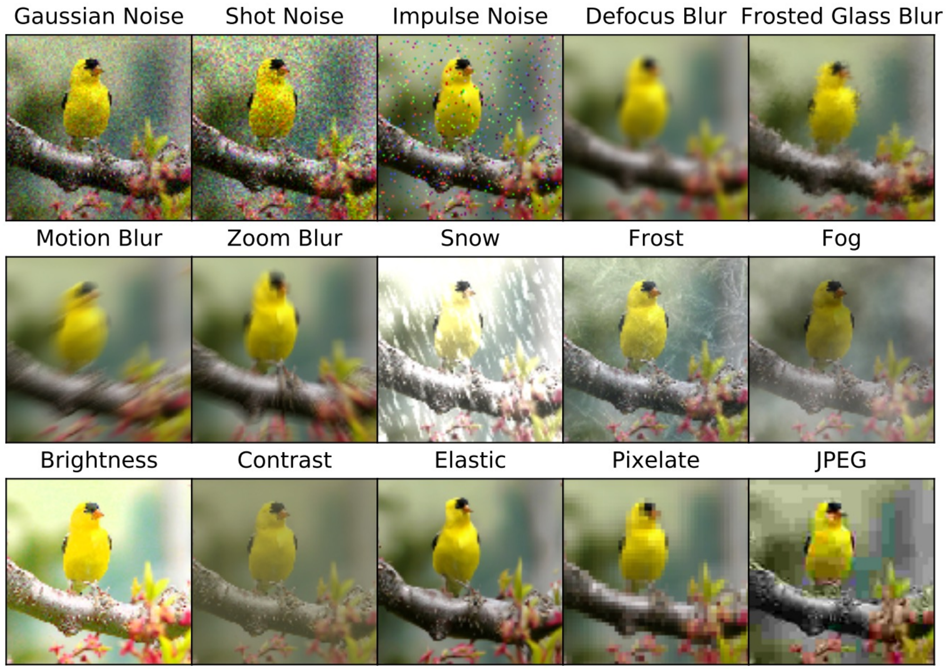
(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

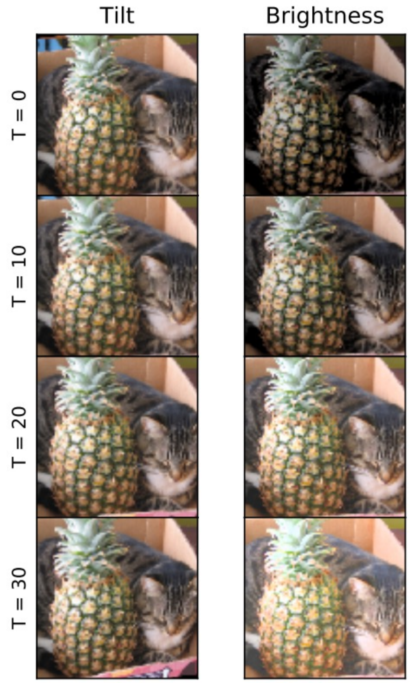
(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan



Temporary solution: Data Augmentation (**Style Transfer**)
ImageNet -> Stylized-ImageNet

Common Corruptions





ImageNet-C:

- 15 types of noise
- 5 severity levels

ImageNet-P:

- 10 types of perturbation

Current solution: **Data augmentation** vs. **Adversarial Training**

谢谢！